CANADIAN THESES

67423

THESES CANADIENNES

AUTHOR/AUTEUR

HARLEY, D.D.

SIMULATED TAILORED TESTING...

ISBN

0-315-19515-0

CANADIAN THESES ON MICROFICHE

THÈSES CANADIENNES SUR MICROFICHE

National Library of Canada
Collections Development Branch

Canadian Theses on
Microfiche Service

Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada
Direction du développement des collections

Service des thèses canadiennes
sur microfiche

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

Canada

NL 339 (r. 86/09)

0-315-19515-0

National Library
of Canada

Bibliothèque nationale
du Canada

173

Canadian Theses Division

Division des thèses canadiennes

Ottawa, Canada
K1A 0N4

67423

PERMISSION TO MICROFILM — AUTORISATION DE MICROFILMER

- Please print or type — Écrire en lettres moulées ou dactylographier

Full Name of Author — Nom complet de l'auteur

Dwight David Harley

Date of Birth — Date de naissance June 15, 1953	Country of Birth — Lieu de naissance Canada

Permanent Address — Résidence fixe

3311 - 47 nd Street
Edmonton, Alberta
T6L 4R9

Title of Thesis — Titre de la thèse

Simulated Tailored Testing of the CCAT

University — Université

University of Alberta

Degree for which thesis was presented — Grade pour lequel cette thèse fut présentée

Ph. D

Year this degree conferred — Année d'obtention de ce grade 1984	Name of Supervisor — Nom du directeur de thèse Dr. S. Hunka

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

Date April 24nd, 1984	Signature Dwight Harley

NL-91 (4/77)

THE UNIVERSITY OF ALBERTA

SIMULATED TAILORED TESTING OF THE CCAT

by

© DWIGHT DAVID HARLEY

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA
SPRING, 1984

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR Dwight David Harley

TITLE OF THESIS Simulated Tailored Testing of the CCAT

DEGREE FOR WHICH THESIS WAS PRESENTED Doctor of Philosophy

YEAR THIS DEGREE GRANTED Spring 1984

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(SIGNED) Dwight D. Harley

PERMANENT ADDRESS:

3311 - 47 Street
Edmonton, Alberta
Canada T6L 4R9

DATED April 24 1984.

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled Simulated Tailored Testing of the CCAT submitted by Dwight David Harley in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Measurement and Evaluation.

S. Hunka

Supervisor

Thomas O. Maguire

External Examiner

Date April 24, 1984

To my parents.

... statistics are, or at least may be, something beyond tabulation and book-keeping.

E. L. Thorndike, 1913

Abstract

This study attempted to determine the effect, if any, a tailored testing strategy had on the ability estimates obtained from the vebal battery of the Canadian Cognitive Ability Test, level F.

A real data simulation study, based on a sample of 4057 grade 9 students, was conducted to compare results from a tailored testing with those from a conventional administration.

The comparison between the tailored and conventional methods of testing revealed exceedingly strong correlations between corresponding subtest scores. Significant reductions were achieved in the number of items that would be administered. This reduction was shown to have a minimal negative effect on the precision of measurement.

Acknowledgements

I wish to extend my appreciation to the Edmonton Public School Board and Mrs. Anne Mulgrew for their generosity in providing this study with the necessary data. Without their cooperation such an extensive data set would have been difficult and expensive, if not impossible, to obtain.

I would like to express gratitude and appreciation to my research supervisor Dr. Steve Hunka, and to Dr. T. O. Maguire, and Dr. D. Sawada, for their guidance, help, and counsel. Dr. A. Olsen and Dr. R. Traub are also thanked for their efforts and suggestions.

To my wife, family, and friends, my sincere thanks for their understanding and encouragement which were necessary in completing this study.

Finally, thanks to the Division of Educational Research Services and its staff who provided me with the support and academic freedom which allowed me to pursue my goal.

Table of Contents

Chapter Page

1. INTRODUCTION 1
  - 1.1 The Problem 9
  - 1.2 Definition of Terms 11
  - 1.3 Research Questions 13
  - 1.4 Delimination of the Study 13
  - 1.5 Justification 14

2. LITERATURE 18
  - 2.1 The Theory of Latent Traits 18
    - 2.1.1 In General Terms 20
    - 2.1.2 A Latent Trait 22
    - 2.1.3 Assumptions 25
      - 2.1.3.1 Unidimensionality 25
      - 2.1.3.2 Local Independence 26
      - 2.1.3.3 Test Speededness 26
      - 2.1.3.4 Item Characteristic Curves 27
    - 2.1.4 Normal vs Logistic Ogive 27
    - 2.1.5 Logistic Test Models 28
    - 2.1.6 Parameter Estimation 36
    - 2.1.7 Information 41
    - 2.1.8 Applications 44
  - 2.2 Tailored Testing 46
  - 2.3 Canadian Cognitive Abilities Test 60
    - 2.3.1 An Overview 61
    - 2.3.2 The CCAT Test Package 61
    - 2.3.3 Item Difficulty and Discrimination 62

2.3.4 Norming 63
2.3.5 Subtest Intercorrelations 63
2.3.6 Reliability 64
2.3.7 Validity 65
2.3.8 In Summary 66
3. METHOD 67
3.1 The Instrument 67
3.2 Design 68
3.3 Models 69
3.3.1 Response Model 70
3.3.2 Scoring Model 71
3.4 Population and Data Collection 72
3.5 Procedure 73
3.5.1 Item Calibration 73
3.5.1.1 LOGIST 73
3.5.2 Rescoring Conventional Administration 74
3.5.3 Tailored Simulation 74
3.5.3.1 Intra-subtest Branching 75
3.5.3.2 Inter-subtest Branching 77
3.5.3.3 Computer Simulation - SIMUTATER ...80
3.6 Research Objectives 82
4. Results 83
4.1 Preliminary Analysis 83
4.1.1 Item Calibration 83
4.1.2 Subtest Ordering 89
4.1.3 Differential Entry Points 90
4.2 Main Analysis 91

4.2.1 Termination Criteria 91

4.2.2 Correlation Analysis 93

4.2.3 Comparison of Test Length 96

4.2.4 Efficiency Analysis 97

4.3 Evaluation of Inter-Subtest Branching Strategy .123

4.4 Increased Precision 125

4.5 Conclusions 129

5. Discussion 130

5.1 Summary 130

5.2 Discussion and Implications 131

5.3 Directions For Future Study 137

5.4 A Final Word 140

Bibliography 142

6. Appendix A 149

7. Appendix B 154

8. Appendix C 167

List of Tables

Table Page

2.1 Medians and Semi-inter Quantile Ranges of Item Difficulty and Item Discrimination.............63

2.2 Inter-subtest Correlations...........................64

2.3 Correlations Between the CTBS Subtests and the CCAT Verbal Battery..........................65

4.1 Item Parameters for Subtest A.......................84

4.2 Item Parameters for Subtest B.......................85

4.3 Item Parameters for Subtest C.......................86

4.4 Item Parameters for Subtest D.......................87

4.5 Final Subtest Sizes.................................89

4.6 Observed Inter-subtest Correlations.................90

4.7 Correlations Between Raw Scores and Tailored Testing Estimates of Ability...............94

4.8 Correlations Between Conventional and Tailored Testing Estimates of Ability...............95

4.9 Mean and Standard Deviations of the Number of Items Administered.........................96

4.10 Mean, Standard Deviation, and Maximum Deviation of Tailored Testing Ability Estimates From C_{00}..............................122

4.11 Mean and Standard Deviations of the Number of Items Administered Without an Inter-Subtest Branching Strategy..................125

4.12 Maximum, Minimum, and Range Values of Ability Estimates................................126

4.13 Differences Between Ability Estimates of Two Subjects at the Lower End of the Ability Scale....................................128

4.14 Differences Between Ability Estimates of Two Subjects at the Upper End of the Ability Scale....................................128

6.1 Extended Item Parameters for Subtest A............150

Table Page

6.2 Extended Item Parameters for Subtest B............151

6.3 Extended Item Parameters for Subtest C............152

6.4 Extended Item Parameters for Subtest D............153

7.1 Mean Number of Items and Posterior Variances for Subtest 1 Under Criteria Levels 0.10 and 0.05..................................155

7.2 Mean Number of Items and Posterior Variances for Subtest 1 Under Criteria Levels 0.025 and 0.01.................................156

7.3 Mean Number of Items and Posterior Variances for Subtest 1 Under Criteria Levels 0.001 and 0.0..................................157

7.4 Mean Number of Items and Posterior Variances for Subtest 2 Under Criteria Levels 0.10 and 0.05..................................158

7.5 Mean Number of Items and Posterior Variances for Subtest 2 Under Criteria Levels 0.025 and 0.01.................................159

7.6 Mean Number of Items and Posterior Variances for Subtest 2 Under Criteria Levels 0.001 and 0.0..................................160

7.7 Mean Number of Items and Posterior Variances for Subtest 3 Under Criteria Levels 0.10 and 0.05..................................161

7.8 Mean Number of Items and Posterior Variances for Subtest 3 Under Criteria Levels 0.025 and 0.01.................................162

7.9 Mean Number of Items and Posterior Variances for Subtest 3 Under Criteria Levels 0.001 and 0.0..................................163

7.10 Mean Number of Items and Posterior Variances for Subtest 4 Under Criteria Levels 0.10 and 0.05..................................164

7.11 Mean Number of Items and Posterior Variances for Subtest 4 Under Criteria Levels 0.025 and 0.01.................................165

Table Page

7.12 Mean Number of Items and Posterior Variances for Subtest 4 Under Criteria Levels 0.001 and 0.0 166

8.1 Observed Efficiencies For Subtest 1 168

8.2 Observed Efficiencies For Subtest 2 169

8.3 Observed Efficiencies For Subtest 3 170

8.4 Observed Efficiencies For Subtest 4 171

List of Figures

Figure Page

2.1 Normal and Logistic Ogives.....29

2.2 Three-parameter Logistic Test Model.....31

2.3 Two-parameter Logistic Test Model.....33

2.4 One-parameter Logistic Test Model.....34

2.5 Item Information Curves, $I(\Theta,u)$.....43

2.6 Test Information Curve, $I(\Theta)$.....45

3.1 Branching Strategies.....81

4.1 Theoretical Test Information Curves.....98

4.2 Averaged Conventional Test Information Curves.....99

4.3 Averaged Test Information Curves For Subtest 1.....100

4.4 Averaged Test Information Curves For Subtest 2.....101

4.5 Averaged Test Information Curves For Subtest 3.....102

4.6 Averaged Test Information Curves For Subtest 4.....103

4.7 Mean Posterior Variance vs Estimated Theta For Subtest 1.....106

4.8 Mean Posterior Variance vs Estimated Theta For Subtest 2.....107

4.9 Mean Posterior Variance vs Estimated Theta For Subtest 3.....108

4.10 Mean Posterior Variance vs Estimated Theta For Subtest 4.....109

4.11 Mean Number of Items vs Estimated Theta For Subtest 1.....110

4.12 Mean Number of Items vs Estimated Theta For Subtest 2.....111

Figure Page

4.13 Mean Number of Items vs Estimated Theta For Subtest 3....................................112

4.14 Mean Number of Items vs Estimated Theta For Subtest 4....................................113

4.15 Observed Efficiencies For Subtest 1...............115

4.16 Observed Efficiencies For Subtest 2...............116

4.17 Observed Efficiencies For Subtest 3...............117

4.18 Observed Efficiencies For Subtest 4...............118

1. INTRODUCTION

The past few years have been witness to a revival of conservative attitudes in educational philosophy. With free-schools, failure-free schools, and the "do-your-own-thing" rhetoric of the sixties and early seventies passing on into time, words such as assessment, accountability, evaluation, and standardization are once more finding their way into the educational forum. Practitioners as well as theoreticians are becoming affected by this shift in the educational climate. Large scale testing and evaluation programs, such as the National Assessment project in the United States, and many smaller provincial or district programs are now being established to begin assessing the present status of our educational systems. This change has brought the often maligned but seldom understood area of measurement once again into the forefront of educational research and reform.

Much has been written in attempting to define the term "measurement." Three decades ago Stevens (1951) interpreted measurement as "the assignment of numerals to objects or events according to rules" (p. 1). Since then, this definition of measurement has become popular and frequently appears in several texts dealing with behavioural methodology (eg. Kerlinger, 1973).

The "assignment of numerals" referred to by Stevens facilitates empirical description of the quantitative and qualitative characteristics of observable phenomena. The assigned numerals innately lack meaning and serve only as symbolic labels until such time as quantitative associations, which ultimately transform these symbols into numbers, are made. The rules used to form the numerical relationships are systematic procedures which define correspondences between observable entities and quantitative representation. The construction and application of these rules of correspondence are germane to the art of measurement. Measurement can thus be thought of as the study of procedures and techniques used in the development and implementation of instruments that facilitate mappings of observable phenomena onto the real number system.

The oldest and most commonly used instrument of mapping employed in the area of educational and psychological measurement is the written test. Generically, a test is thought to be any systematic procedure for eliciting responses characteristic of the attribute or trait in question (Sax, 1974).

Records (Miyazaki, 1963/1976; Ebel, 1972) have shown that the earliest use of written tests was by the ancient Chinese civil service in about the year 2300 B.C. Potential employees were assessed by means of competitive examinations. Chinese governments maintained this practice until the turn of the present century.

In the Western World, universities of the Medieval and Renaissance periods relied exclusively upon the use of oral examinations. It was not until mid-way through the sixteenth century, when paper became more plentiful, that written tests began to replace their oral cousins. The Jesuits are generally considered the first in recent times to have frequently used written examinations for evaluative and placement purposes.

In our society, measurement for the purposes of educational and psychological decision making had its beginnings in the early 1900's from contributions by people such as Spearman (1914), Binet (1916), and Thorndike (1919). From these works evolved classical test score theory and its related statistical framework.

One of the most useful developments yielded by classical test score theory is the weak form of the true-score model. This model is a mathematical function which defines a relation between the hypothesized true test scores and the observed test scores based upon a set of relatively weak assumptions. These assumptions and hence the model are regarded as weak in the sense that they are considered to be "obviously satisfied by most data" (Lord & Novick, 1968, p. 25). Because of this characteristic and because it has allowed us to do a lot of useful things, the weak true-score model has gained widespread acceptance as the standard testing model of educators and evaluators.

However, despite its popularity, the weak true-score model has not become the panacea of psychometricians. On the contrary, Hambleton (1979, 1980) has discussed the many limitations of this model and related item technology (*i.e.* conventional item analysis). He has suggested that one of its most severe limitations is that the values of the two common item indicies, item difficulty and item discrimination, are dependent upon the group ability characteristics of the examinee-sample being tested. This limits the use of obtained item statistics to the construction of tests that will be administered to groups similar in nature to that of the sample from which the item statistics were initially calculated. Thus item parameters that remain invariant across groups would be a definite advantage to test designers and psychometricians for purposes of test construction.

A further deficiency of the classical model is the dependence of examinee ability estimates upon the particular choice of items in a test. Without going into the complexities of test equating (in the conventional sense), we note that inter-examinee comparisons of ability are not possible unless all examinees have responded to identical or statistically parallel sets of items. In today's large, competitive school systems where district and system wide comparison of students is often desirable, it is impractical to mount large scale testing campaigns that require all students to respond to the same testing instrument; this approach to testing is too expensive for the yield of information it provides for program evaluation. Hence the ability to compare examinees not responding to the same sets of items would be of great value to our educational systems.

These limitations provided the motivation for psychometricians to search for more appropriate testing models. One group of such models that has received a great deal of consideration is the group of latent trait models. The theory of latent traits has been referred to as item-characteristic curve theory, item-response theory (IRT), and also modern test theory. This class of models is a relatively recent development and can be considered to have had its origins with Lord's classic work *A Theory of Test Scores* (1952).

The plausability of latent traits is established in detail in section 2.1.2, but generally a latent trait (or latent ability as it is sometimes called) can be thought of as some unobservable characteristic of human performance which ultimately is identified and defined in terms of observable variables. Latent trait models are formalized mathematical functions that define relationships between observable variables and the latent trait. Such models presume that quantification of latent traits can be used to predict an examinee's performance in related situations (Lord & Novick, 1968). In other words, latent trait models provide a method of inferring estimates of an examinee's latent ability from a set of test items.

One of the attractions of latent trait theory is that it overcomes the disadvantages of conventional item analysis discussed above. Given a large sample of examinees, item parameters are independent of the group on which they were calculated (*i.e.* parameter invariance) and inter-examinee comparisons are facilitated since examinee ability estimates are not dependent upon the choice of test items that comprise the measuring instrument. A particularly nice feature is that mathematical expressions relating item performance to examinee ability level are inherent in the model. For these reasons and others, latent trait theory has become, as Hambleton (1980) puts it, "the hottest topic in the measurement field."

When Lord published his treatise in 1952, his theory preceeded practical application by more than two decades. Hambleton and Cook (1977) advanced the following five reasons as to why IRT took so long to become popular:

1. Potential users were discouraged by the mathematical complexity of IRT.
2. The majority of the work was aimed toward theoreticians rather than practioners.
3. Efficient, high speed computer programs [and in the early years of IRT, efficient, high speed computers], were not available for item parameter estimation.
4. Skeptics questioned the advantages gained through IRT.
5. Concerns were expressed as to the tenability of the model's strong assumptions.

As time passed the problems of tractability were solved through the work of people such as Birnbaum (1968) and Wood, Lord, and Wingersky (1976). Further work by Lord (1968) and others began to silence the skeptics by demonstrating IRT's utility, while special journal issues (eg. *Journal of Educational Measurement*, 14(2), 1977), conference sessions and workshops, and a myriad of readable journal articles brought the theory of latent traits to field practioners. The late seventies saw IRT getting the recognition it deserved.

With the realization of Green's (1970, p. 194) prophecy of "the inevitable computer conquest of testing," the theory of latent traits has proven itself to be useful as well as practical. A most natural blend of item response theory and computer technology is the basis of tailored or adaptive testing. Tailored testing is claimed by some researchers (Urry, 1977) to have tremendous potential for improving the measurement of psychological traits and abilities.

Tailored testing is a measurement technique in which the test, as the name suggests, is tailored (or adapted) to meet the specific ability level of each examinee. A tailored test is constructed from a pool of precalibrated items, such that the items selected are of approximately average difficulty for a given examinee at a given ability level. Then by implication, each examinee need not respond to the same set of items drawn from the same item pool. It can be shown that administration of items that correspond to the examinee's ability level increases the precision of measurement on an individual basis. For much of the ability range, tailored testing can result in more precise measurements than are obtainable using conventional methods of testing. (This assumes, of course, tests of comparable length.) The difference in precision is dramatic for examinees at ability levels located at either extreme of the ability range.

Lord (1970) considers item economy another possible advantage of tailored testing. A tailored test can attain a level of precision at least comparable to a conventional test even though most examinees will take significantly fewer items. This has two particularly attractive features: the first being item conservation when the item supply is limited, a second being a reduction in the amount of time consumed by testing.

The foregoing model of tailored testing may seem to rest on an unsupportable assumption. It presupposes that the examinee's ability level is known prior to testing, yet the purpose of the test is to measure this level. This problem is solved by starting from a crude (initial) estimate of the examinee's ability level. This estimate may be obtained in any number of ways - through routing tests, estimated from previous data or test results, or through any other means available to the tester. This estimate is used to select and administer an item which is scored and used as a basis to refine the initial estimate. The refined estimate is used for a subsequent item selection and the procedure is repeated until a predetermined termination criterion has been met.

Many variations of this approach exist but they all attempt to match the difficulty of the items administered to the ability level of the examinee.

The above description should make clear that tailored testing is a dynamic process unlike the static methods of conventional testing. Due to its dynamic nature the most practical paradigm for tailored testing is an interactive computer-based model. During the testing procedure an examinee is seated in front of a video display computer terminal on which the test items are presented. The examinee responds to each item by means of a typewriter-like keyboard. The computer then performs the item scoring and subsequent ability re-estimation procedures. If the new estimate is greater than the previous estimate the next item selected will be slightly more difficult, and conversely if the new estimate of ability is lower, then the succeeding item will be slightly easier. The examinee's ability estimate becomes more accurate with each iteration of the testing procedure. The strategy followed is somewhat similar to that of a binary-search algorithm.

1.1 The Problem

With tailored testing's considerable advantages over conventional methods of measurement, it has the potential of becoming one of the most useful tools educators and psychologists have at their disposal. Urry (1977, p. 184) points out that "if tailored testing is to have immediate application, it must use existing test items." This statement suggests the possibility of modifying existing tests for use as tailored tests. If tailored versions of existing tests can produce results equivalent to those obtainable from conventional administrations, then the testing community can begin reaping the benefits of tailored testing without waiting for test developers to construct tests especially designed for adaptive testing. With this as the motivating factor, the purpose of the present study was to determine whether the application of a tailoring strategy to an existing, intact test would yield results comparable to those produced by conventional administrations of the same test.

Clearly, a problem of this magnitude cannot be answered by just one study. Considering the vast number of tests currently available, it was necessary to reduce the problem to one of manageable proportion, by selecting one test battery and examining the effect a tailoring procedure had on the results of that particular instrument. Thus, this study attempted to determine the effect, if any, a tailoring strategy had on the ability estimates obtained from the verbal battery of the Canadian Cognitive Ability Test, level-F (Thorndike & Hagen, 1974). A simulation study was conducted to compare results from an adaptive testing with those from a conventional administration.

Some previous work had been done in this area (Brown & Weiss, 1977 ; Bejar, Weiss & Gialluca, 1977; Gialluca &

Weiss, 1979), but the instruments employed were of the classroom achievement variety. Note that the tests used in the Bejar et al. and the Weiss and Gialluca studies were specifically designed for tailored testing. Other work (Lord, 1975) was based upon item pools formed by aggregating several, different testing instruments, while still others restricted themselves to measuring a single area (*i.e.* dimension) of content.

The test selected for use in this study was chosen from the domain of standardized mental ability tests, which, it can be argued, are totally different in nature from classroom achievement tests (Bejar, Weiss, & Kingsbury, 1977). The items which formed the item pools are found in the verbal battery of the Canadian Cognitive Abilities Test, level F. A multi-dimensional tailoring strategy was used to obtain ability estimates for each of the battery's hypothesized dimensions.

1.2 Definition of Terms

The following is a collection of terms accompanied by definitions which indicate the precise sense in which these terms were used in this study.

1. administration: a procedure used to obtain ability estimates. In this study the following two procedures were employed:
    a. conventional: the administration of the Canadian

Cognitive Abilities Test in the traditional pencil and paper format.

   b. tailored: a computerized simulative administration of the Canadian Cognitive Abilities Test performed on the basis of raw item data previously collected in the traditional manner.

2. precision (of measurement): the accuracy of measurement in terms of maximum information or equivalently, minimum standard error of estimate.

3. subtest scores: the scores obtained on each of the four subtests of the verbal battery of the Canadian Cognitive Abilities Test. Three sets of subscores were used in this study.
   a. raw scores: the number-correct score for each of the four subtests from the conventional administration.
   b. conventional scores: a latent ability score computed by Owen's (1975) Bayesian scoring procedure for each of the four subtests on the basis of the conventional administration.
   c. tailored scores: a latent ability score computed by Owen's Bayesian scoring procedure for each of the four subtests on the basis of the tailored administration.

1.3 Research Questions

Given that the objective of this study was to assess the effect of the imposition of a tailoring strategy upon a conventional test (*i.e.* Canadian Cognitive Abilities Test, level F, verbal battery) three research questions seem relevant:

1. What are the correlations between the respective subtest scores of the simulated tailored and conventional administrations?
2. Can the tailored testing strategy result in a reduction of the number of items administered while maintaining an acceptable level of precision?
3. Is there a high correspondence between the information curves yielded by the two administrations?

If a high correlation between corresponding pairs of subtest scores could be established, then it was hoped to show that the number of items required to obtain the same precision of measurement under a tailoring strategy would be substantially smaller.

1.4 Delimination of the Study

The study was carried out using item level data collected by the Edmonton Public School Board in a district wide administration of the verbal battery of the Canadian

Cognitive Abilities Test, level F. Item responses from 4057 grade nine students were provided for the study.

1.5 Justification

Research to date has generally established the credibility of tailored testing as a possible alternative to conventional testing methods. However, the majority of previous research has been based upon item pools specifically designed for tailored testing. This study was an attempt to examine the validity of Urry's (1977, p.184) claim that "if tailored testing is to have immediate application, it must use existing test items." Developing item pools specifically designed for tailored tests that measure the same dimensions as currently existing conventional tests would be redundant as well as difficult and expensive - both in terms of time and money. If it is found that already existing test item pools (*i.e.* intact tests or subtests) can be used successfully with tailored testing, then this study will encourage practioners to apply tailored testing to their own testing needs. The study was intended to stimulate immediate application of tailored testing rather than to perpetuate the assumption that tailored testing requires specially designed and complex item pools.

Provided it can be demonstrated that comparable results can be obtained from tailored versions of conventional tests, evaluation as a whole is more likely to be transformed in a way that will make it more efficient. As previously noted tailoring carries the promise of greater precision of measurement for the administration of a smaller number of items than conventional testing. Hence tailored testing promises to be a much more cost effective procedure than conventional testing.

Other advantages of tailored testing are by-products of its computerized format. Response errors are eliminated by having examinees enter their responses directly into the computer rather than on answer sheets which must then be transcribed into machine readable form. The conventional problems of test security are reduced as the need for multiple copies of examination materials is eliminated. Different security risks are created as test security becomes confounded with the problem of computer security. These problems will be subtle as security violations will be more difficult to detect.

Computerized tailored testing eliminates the examiner effect as well as problems involved with the handling and administration of testing materials. The efficiency of the testing environment will be much improved through computerization.

This improved efficiency will not come without cost, however, as different environmental effects must be dealt with. Research (Hedl, O'Neil, and Hansen, 1971; Harley, 1979) on the use of computers as a testing medium has been conducted, but more is needed to substantiate results and assess other effects of computerized measurement.

One area to benefit from tailored testing is computer assisted instruction (CAI). Having both the instructional and evaluative aspects of the educational process located within the same medium should attract more educators to the possibilities of CAI and result in a better selection of instructional strategies.

Computerized testing offers other advantages besides those of a logistical nature. Testing through this electronic medium facilitates the collection of information unobtainable through the more conventional, pencil and paper modes of testing. For example, response latency (*i.e.* the amount of time taken to answer each item) would be very impractical or even impossible to quantify in conventional settings but a trivial task in computerized environments. Access to this kind of information will facilitate further research in the area of response style.

Measurement has progressed a long way since man first became interested in the assessment of mental abilities. From the ancient Chinese civil service exams, to pencil and paper exams, to Spearman's factor analysis, and on to Lord's latent trait theory and all of its applications, the development of measurement has closely paralleled civilization's progression from manual calculating through electronic computation. With the increasing influx of computer technology into the areas of education and psychology, the possibilities for improved measurement and evaluation techniques are limited only by imagination. Tailored testing is one of the more promising of these possibilities and with the recent popularity of latent trait theory, it is hoped that more research will be done to help in the attainment of its full potential.

The following chapter examines the work done in the area of latent trait and tailored testing. A critique of the verbal battery of the Canadian Cognitive Abilities Test, level F is also included.

2. LITERATURE

Over the past few years the family of latent trait testing models has attracted considerable attention from the psychometric and measurement communities. The theory of latent traits is becoming increasingly more important and is forecast by some (Hambleton & others, 1978) to become the method of the future for the measurement of mental abilities. The purpose of this chapter is to provide a relatively brief overview of latent trait theory and then focus upon one of the more promising applications of latent trait theory, tailored testing and its relationship to the problem at hand. The chapter will conclude with a critique of the Canadian Cognitive Abilities Test (Thorndike & Hagen, 1974).

2.1 The Theory of Latent Traits

Thomas Warm(1978) introduced the first chapter of his Primer of Item Response Theory with a most zealous appreciation of its value.

> Item response theory (IRT) is the most significant development in psychometrics in many years. It is, perhaps, what Einstein's relativity theory is to physics. I do not doubt that during the next decade it will sweep the field of psychometrics. It has been said that IRT allows one to answer any question about an item ... a test, or an examinee, that one is entitled to ask. (p. 11)

A comparison of IRT to Einstein's relativity is, perhaps, an overstatement of its importance but it does serve as testimony to the excitement and enthusiasm that many researchers and practitioners have for this field.

Despite Warm's exaggerated views of the value of IRT, some researchers (Weiss & Davison, 1981) find themselves in general agreement with Warm. In their review of the research literature pertaining to test theory for the period 1975 through 1980, they state that research had "concentrated on relatively unimportant developments in reliability theory ... during this period." The review concludes that little progress was made in the advance of classical test theory and that the most significant contributions were made in the development of alternative testing models. These models, which include latent trait theory, were proposed to handle a number of problems which classical test theory has been unable to solve adequately. During the late 1970's IRT was the subject of a considerable amount of research and underwent a period of rapid growth.

Traub and Wolfe (1981), who appeared slightly reserved in their regard for latent trait theory, note that there is still room for substantial growth in many directions. They feel that much research needs to be done before IRT results in a general improvement of testing procedures. A warning was sounded with regard to blind application of IRT models and it was strongly suggested that consideration of the models and their fundamental assumptions needs to be given prior to application.

> We have tried to bring the reader a message using the jargon of the market place, a message that is applicable to latent trait theory when used to assess educational achievement: *Caveat emptor!* (p. 384)

2.1.1 In General Terms

It is generally considered that Lord's doctoral dissertation *A Theory of Test Scores* (1952) was the birth of latent trait theory or item characteristic curve theory as he called it. The roots of latent trait theory can be traced as far back as the beginning of the twentieth century with Binet, but Lord's dissertation is thought to be the first appearance of IRT in its own right.

Lord (1952) outlined a mathematical model defining the relationship between an obtained test score and the underlying trait (*i.e.* latent trait) or ability being measured. His theory was expressed as a function of the difficulties and discriminating powers of the individual test items. The theory was developed for application to tests composed of free-response items but it was argued that extension to item types that may be answered correctly by guessing (eg. multiple choice, true-false, etc.) is mathematically straightforward. Lord sees his theory as an aid in interpreting test scores, as well as a major tool for test design, construction, and analysis. In the preface he claims that this new theory

> is more powerful ... than direct application of the classical theory of errors starting with the broad assumption that test score and "true score" differ by normally distributed, independent errors of measurement. (p. v)

In the following year Lord published two papers (1953a, 1953b) which extended and refined his previous work. Both articles wrestled with the problem of developing a metric for the ability or trait that underlies the the raw score of the test. It was desirable that such a metric be invariant across tests. These two papers along with Lord's dissertation, which were much more theoretical than practical in nature, represented the state of the art with respect to latent trait theory at that time.

One of the many conclusions reached in these papers is of special interest to this study. Quoting from Lord (1953b):

> A given examinee's ability can be estimated more accurately by administering free-response items that are of 50% difficulty for examinees like him than administering free-response items at any other single difficulty level. (p. 75)

This maxim underlies some item selection strategies employed in tailored testing. As will be discussed in a later chapter, this study followed one such strategy.

Not much was done in the area of latent trait theory during the rest of the 1950's and early 1960's, probably due to its complex mathematical nature and a lack of the computer technology (both hardware and software) necessary to estimate model parameters and apply them to practical testing problems.

In 1968, Birnbaum contributed four chapters to the well known textbook by Lord and Novick (1968), in which much of the mathematics involved in the more common latent trait models was worked out in great detail. Treatment was given to both the two and three parameter forms of the logistic and normal ogive models.

Since then a large number of journal publications have been written and several computer programs and packages have been developed. Wright and Stone (1979) released a textbook dealing with the one parameter Rasch model and Lord (1980) has recently written another textbook directed toward practical applications of IRT. The field of latent trait theory has definitely opened up to those people looking for reasonable alternatives to the classical test theory.

2.1.2 A Latent Trait

Up to this point the theory of latent traits has been discussed without stating what a latent trait is. The notion of a latent trait must be defined before detailed theoretical development can be provided.

Lord (1968) refers to any defined human characteristic accounting for some facet of an individual's behaviour as a "trait." Examples of such traits include mathematical ability, verbal ability, physical conditioning, and creativity.

The adjective "latent" connotes a characteristic that is underlying, hidden, or unobservable. Hence, a "latent trait" is thought to be an unobservable human characteristic, a characteristic without any direct physical referent. Such traits are inferred from observable or manifest variables, with their identification being the subject of considerable theoretical model building and testing.

There are basically two major (and perhaps polar) views concerning the nature of latent traits. The first view considers a latent trait as a hypothetical construct, which exists independently and apart from the items that make up the test. This view assumes that the items elicit latent responses, which are translated into observed responses to the items. Hypothetical constructs require justification on psychological grounds; providing this justification is the objective of construct validation studies.

The alternative is the empirical view with latent traits regarded as manifestations of inter-item and item-response relationships. Those who take this stance deem model fit the important concern, not the validity of a construct. As this view stresses inter-item relationships, perhaps it is more deserving of the name "item response theory" than "latent trait theory."

An important assumption to latent trait theory is that the latent continuum is unidimensional. Given that verbal ability is a construct, would it be reasonable to assume that it is also unidimensional? - perhaps not. Therefore, applying latent trait theory under the first view could violate an important assumption of the model. The empirical view avoids this concern by viewing IRT models as formalizations of the relationship between items and responses, and ignores the question of the meaning of the continuum. This does not seem to be totally satisfactory, as traits or abilities are defined in terms of items and item responses rather than theory.

The approach taken in the present study was one of compromise. A latent trait was considered to be a one dimensional component of a hypothetical construct. After all, we know that constructs such as verbal and arithmetical ability are well defined in the educational literature. Furthermore, it was assumed that the test items measured the construct and that the item-response relationships were characterized by the latent trait model. The view was taken that both the meaning of the continuum as well as the response model are important in defining a latent trait.

The ultimate objective of latent trait theory is to quantify an individual's position on a trait in terms of a continuum known generally as an ability level. In this study, the symbol Θ (theta) is used to represent ability level on the trait being measured. The ability continuum, thought to be an interval scale, theoretically ranges from minus infinity through positive infinity (*i.e.* $-\infty<\Theta<\infty$).

2.1.3 Assumptions

Certain initial assumptions are fundamental to the development of most mathematical models and theories. The following discussion addresses four basic assumptions germane to the theory of latent traits.

2.1.3.1 Unidimensionality

A basic assumption of latent trait theory is that the dimensionality of the latent space (*i.e.* space defined by a set of mutually orthogonal traits) is one which implies that a single trait or ability underlies test performance. Multidimensional latent trait models do exist but Lord (1980) feels that their practical application is beyond the present state of the art. Hambleton (1979) suggests that the assumption of unidimensionality simplifies the interpretation of the resulting scores.

Methods that attempt to assess the dimensionality of a set of items are plentiful and diverse. Extensive collections of such methods are reviewed by Hattie (1983(a), 1983(b)) but as reported in both papers "there are still no known satisfactory indices."

2.1.3.2 Local Independence

It is assumed that an examinee's responses to all items are independent, or in other words, the probability of answering a given item correctly is unaffected by responses to the other items. An especially nice consequence of this assumption is that the probability of a given response vector (*i.e.* series of responses) is just the simple product of the probabilities of the individual item responses.

Local independence is considered by Lord (1980) to be a natural consequence of the assumption of unidimensionality and not necessarily an additional assumption. Unidimensionality is a sufficient condition for local independence however the converse is not true.

2.1.3.3 Test Speededness

Latent trait theory assumes a power test rather than a speeded test. Speeded tests violate the assumption of local independence since the later items of a test may be classed as wrong simply because the examinee failed to reach that point in the test. In IRT it is assumed that an examinee attempts each item. For speeded or pure speed tests Lord and Novick (1968) recommend other testing models such as the Poisson or gamma models.

2.1.3.4 Item Characteristic Curves

The fourth basic consideration of latent trait theory is that all item characteristic curves (ICCs) take the general shape of the normal ogive. The ICC is a function unique to each item that relates the probability of a correct response to the ability level (Θ). This probability depends only on the shape of the ICC and is independent of the distribution of Θ.

Statistically the ICC represents the nonlinear regression of item score on ability. Then, by definition[1], item characteristic curves are invariant up to a linear transformation of the scale for ability, across groups of examinees (Lord & Novick, 1968). This is one of the most important and useful characteristics of latent trait theory. Figures 2.2 through 2.4 provide several examples of item characteristic curves.

2.1.4 Normal vs Logistic Ogive

In the previous sub-section, in which item characteristic curves were discussed, the assumption was made that all ICCs take the general shape of the normal ogive. However, since the normal ogive is a transcendental function[2] the mathematics encountered in working with it is awkward, and a simpler model would be preferred.

---

[1] Regression functions are independent of the frequency distribution of the predictor variable (Lord, 1980).
[2] A curve not representable by an algebraic function.

An algebraic function that is mathematically simple and that can be made to yield a very close approximation to the normal ogive is the logistic ogive (equation 1) below.

$$\Psi(x)=\exp(x)/(1+\exp(x))=1/(1+\exp(-x)) \quad (1)$$

Haley (cited in Birnbaum, 1968) illustrated the closeness of this approximation by determining that

$$|\Phi(x)-\Psi[(1.7)x]|<0.01 \quad \text{for all } x.$$

In simple terms this means that $\Psi[(1.7)x]$ (logistic ogive) differs from $\Phi(x)$ (normal ogive) by less than 0.01, uniformly for all x. The closeness of the two functions is illustrated graphically in Figure 2.1. Hence, without loss of interpretability, the logistic ogive can be substituted as a mathematically tractable alternative to the normal ogive, the result being a class of models known as the logistic test models (Birnbaum, 1968).

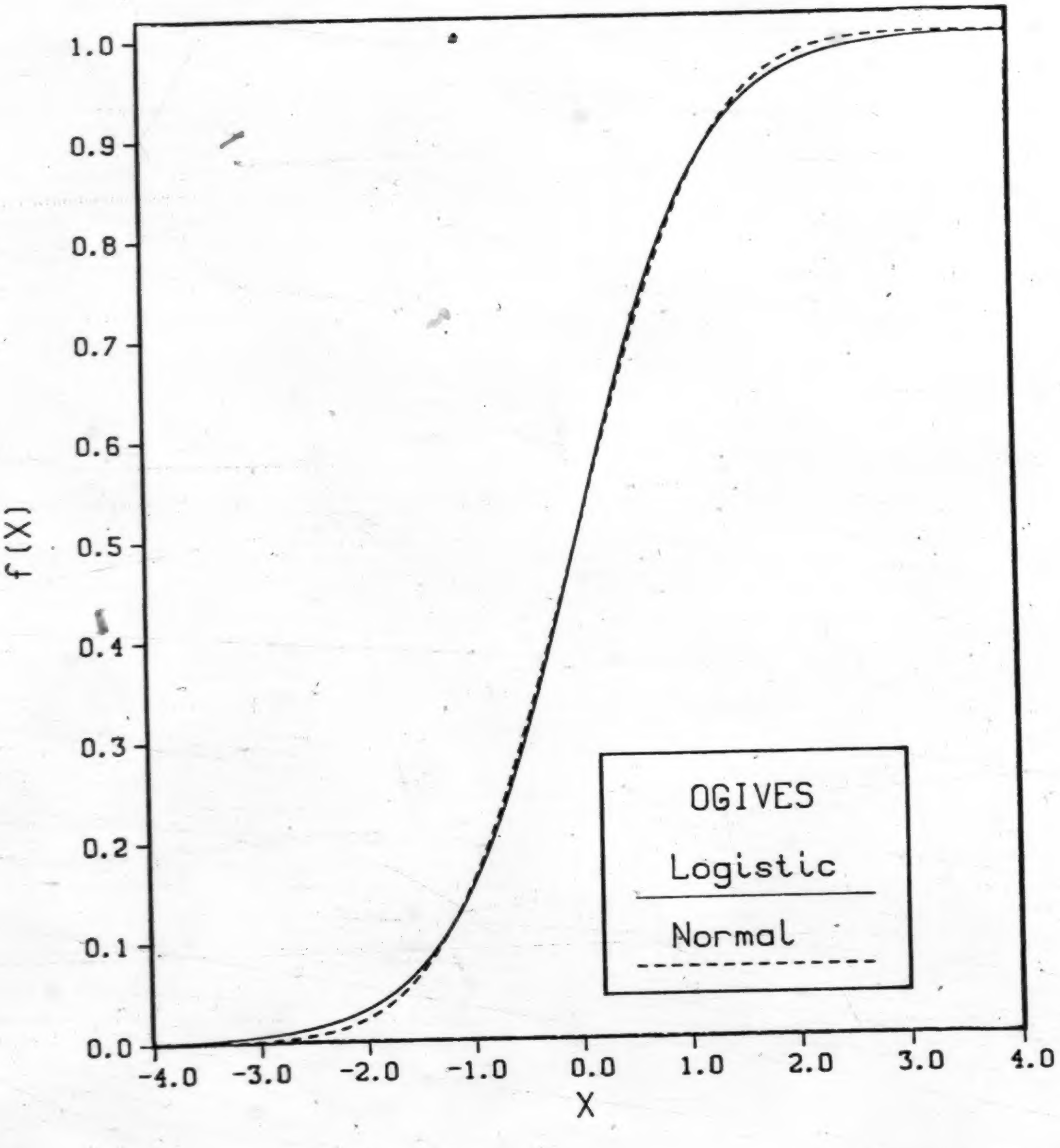
2.1.5 Logistic Test Models

There are three common varieties of the basic logistic test model. Each variety is characterized by the specific attributes of their associated item characteristic curves which result directly from initial assumptions about the item pool.

Perhaps the most popular member of the logistic test model family is Birnbaum's (1968) three-parameter logistic model. Its popularity is due to claims (Urry, 1971; Warm, 1978) that it best describes the real world in terms of responses to multiple choice items. The present study will

Figure 2.1

Normal and Logistic Ogives



be based upon this model.

The general form of this model is given in equation 2:

$$P(1|\Theta)=c+(1-c)[1+\exp(-1.7a(\Theta-b))]^{-1} \quad (2)$$

where:

a is the discrimination index,

b is the difficulty level, and

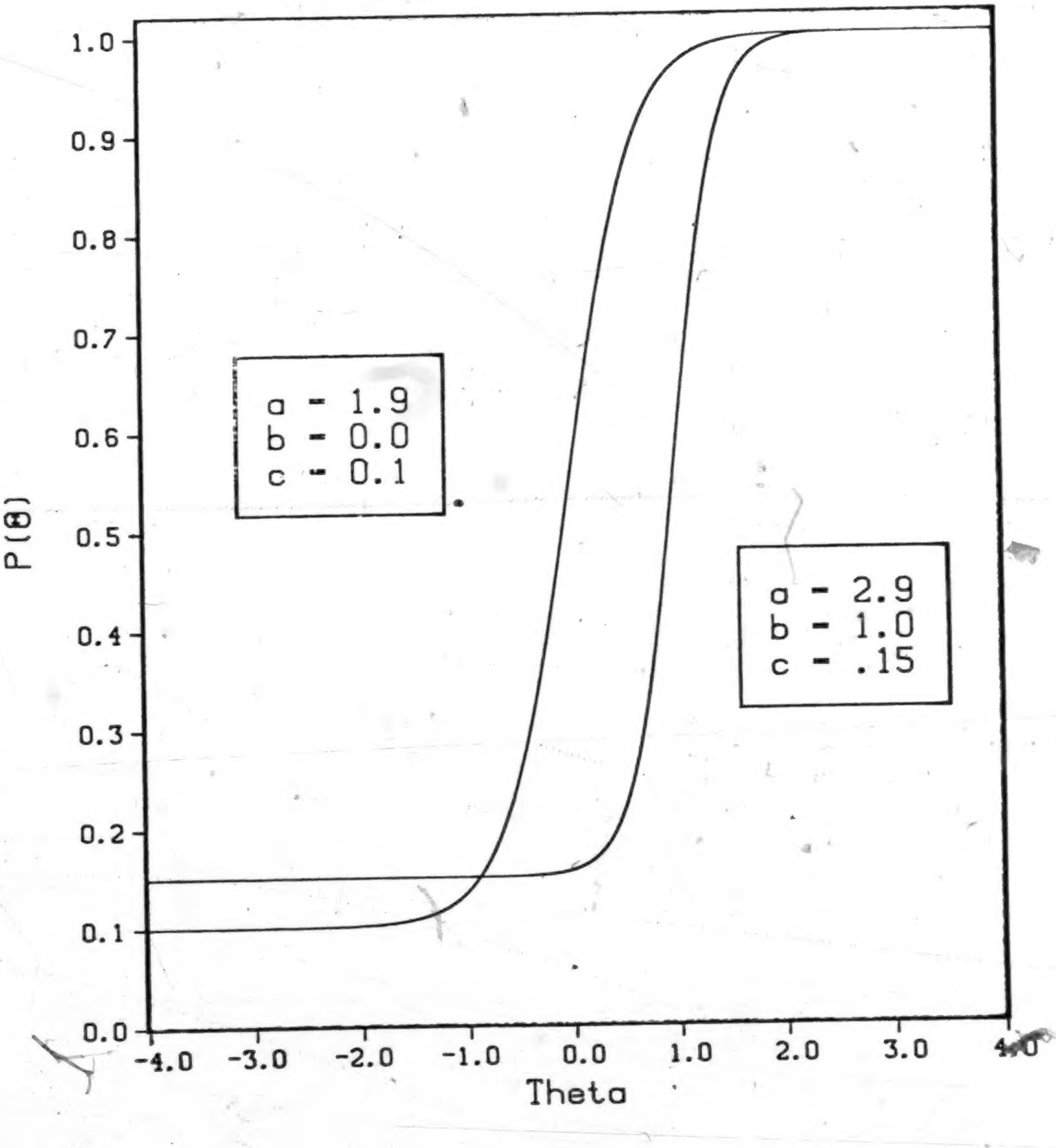
c is the pseudo-chance level for a given item.

Here it is assumed that all items are of varying difficulty level and discriminating power. Also the items are free to vary on a guessing or pseudo-chance parameter. Under classical theory a true guessing or chance parameter would be equal to the inverse of the number of distractors belonging to the item. Such a value is not realistic because examinees seldom guess randomly when a response is not known. Two examples of ICCs associated with this model are provided in Figure 2.2.

A special case of this model is Birnbaum's (1968) two-parameter logistic model. It assumes a pseudo-chance paramater of zero (*i.e.* c=0) for every item but as in the three-parameter model allows for varying levels of difficulty and discrimination. Since the two-parameter model ignores guessing, it seems a priori that this would not be as useful a model for this study, which involved multiple-choice items, as would the three-parameter model. The two-parameter model is more appropriate for dichotomously scored free-response items than a test composed of multiple choice items. This model is basically

Figure 2.2

Three-parameter Logistic Test Model



the logistic analogue of the model Lord proposed in 1952. Illustrated in Figure 2.3 are two examples of ICCs that correspond to the two-parameter model.

Another model that has recently received a great deal of attention is the one-parameter logistic model, more commonly known as the Rasch[3] model. The Rasch model can be viewed as a special case of the Birnbaum three-parameter model, but it can also be derived in an alternative manner, one that is independent of other latent trait models.

When the Rasch model is derived from Birnbaum's three-parameter model, it must be assumed that all items have equal discriminating power (*i.e.* $a=a'$) and have a zero pseudo-chance level (*i.e.* $c=0$). Items are allowed to vary only in terms of their difficulty. Two Rasch model ICCs are illustrated in Figure 2.4.

Hambleton (1979) discusses several advantages of using the Rasch model. He claims that since the model has fewer parameters it is both simpler to work with and easier to explain. As a result of having only one parameter, that of item difficulty, the problem of parameter estimation is minimized. These are considered to be important features in promoting the use of the Rasch model, as practitioners are thought to prefer less complicated models. In addition to these advantages, Hambleton also notes that (assuming model-data fit) the rank order of examinees on test score remains the same on ability level unlike other latent trait

---

[3] Named after the Danish mathematician Georg Rasch (1966).

Figure 2.3

Two-parameter Logistic Test Model

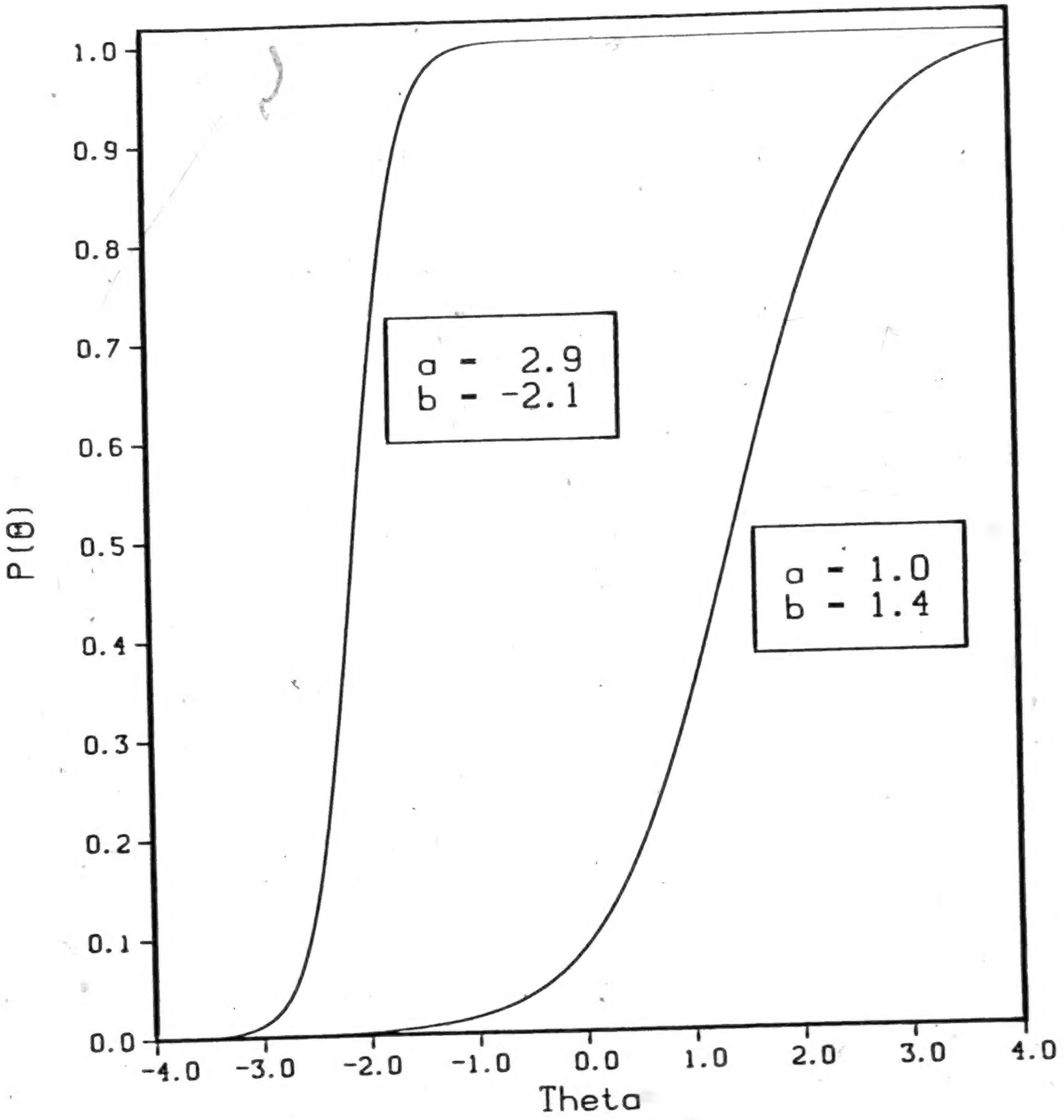
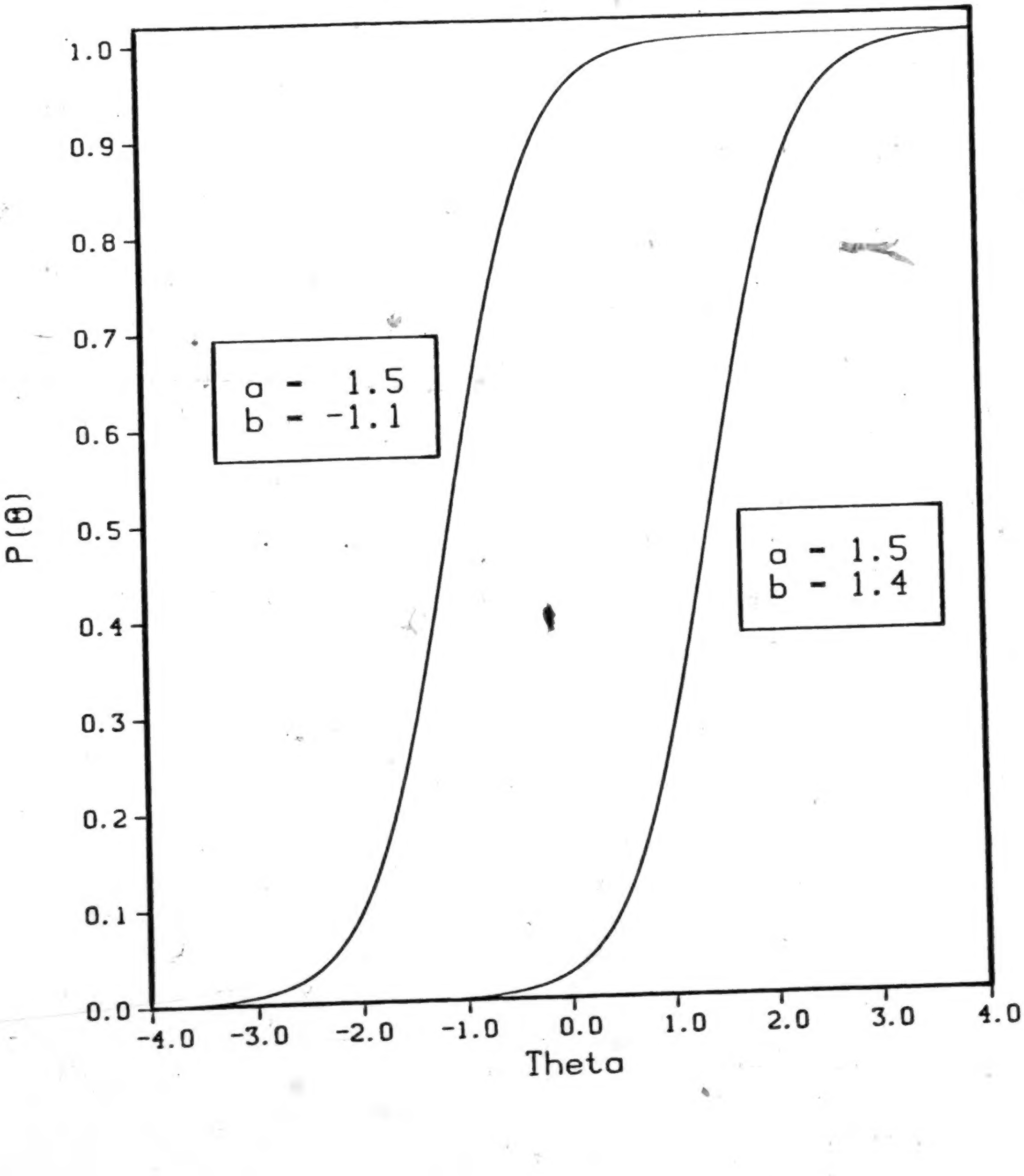


Figure 2.4

One-parameter Logistic Model



models.

The assumptions that make the Rasch model so appealing and simple are actually very restrictive and certainly unrealistic when applied to data from multiple choice tests. Divgi (1981, p. 2) warns that the "use of an unrealistically simple model can lead to serious errors." In a study of the Rasch model for multiple choice items, Divgi (1981) dismisses as inadequate most studies in which the Rasch model has been found to be satisfactory on the grounds that they have lacked powerful inferential techniques. Using a modified test of misfit, one that is substantially more powerful than previous tests, Divgi (p. 1) concluded that "available evidence suggests that the Rasch model cannot fit multiple choice items" and that the three-parameter model is superior for this type of item.

Another problem with the one-parameter Rasch model is its failure to possess a high degree of item parameter invariance. Green and Divgi (1981) showed that of the three popular logistic test models, the Rasch model has the least item parameter invariance while Birnbaum's 3-parameter model has the most. This is indeed a serious consideration when selecting a model, as parameter invariance and objectivity are claimed as major advantages of latent trait theory. For these reasons other authors (Urry, 1977) consider the three-parameter logistic model superior to the Rasch model for purposes of tailored testing. The three-parameter model was used in the present study.

A recently developed model worthy of just a brief mention is the four-parameter logistic model. This model was studied by Barton and Lord (1981) who were motivated by the concern that clerical errors committed by high-ability students on easy (*i.e.* less difficult) items would result in ability estimates being severely under-estimated by the three-parameter model. A fourth parameter, an upper asymptote slightly less than 1, was added to the three-parameter model to alleviate this concern. This model is expressed as follows:

$$P(\Theta|1)=c+(\omega-c)[1+\exp(-1.7a(\Theta-b))]^{-1} \quad (3)$$

where:

a is the discrimination index,

b is the difficulty level,

c is the pseudo-chance level for a given item, and

ω is the upper asymptote.

This model was compared to the three-parameter model on several sets of data and the ability estimates were found to be generally unchanged. Barton and Lord concluded that there was little evidence to encourage the use of this model.

2.1.6 Parameter Estimation

A serious impediment to the use of latent trait models, in particular the three-parameter logistic model, is the problem of parameter estimation. In practice it is necessary to estimate the item parameters a, b, and c for each item of the test and Θ for each person taking the test. Hence for an *n* item test taken by *N* people, $3n+N$ parameters must be estimated.

Much has been written[4] concerning parameter estimation but as yet, nothing simple and elegant has been proposed. Two commonly used methods, both of which were used in this study, are described below.

A popular technique, which requires electronic computers for implementation, is the method of maximum likelihood estimation (Lord, 1980; Hambleton and others, 1978). This method involves an iterative procedure in which estimates of a, b, c, and Θ are calculated in two stages.

The likelihood function of both ability level and item parameters is the joint distribution of all item scores across all examinees formed under the assumption of local independence (Lord, 1981, pp. 56, 59, 179-181). Maximum likelihood estimates (MLE) are found by taking the partial derivative of the natural logarithm (ln) of the likelihood function with respect to each of the four parameters. The derivatives are then set to zero yielding the likelihood equations. The numerical values of the parameters that set these equations to zero are considered to be the MLE of the parameters.

This method is actually estimating two mutually independent sets of parameters: one set relating to items and another set relating to examinees. When the item parameters are fixed or known, Θ for a given individual is determined simply by solving one equation in one unknown. And similarily, given fixed or known Θ, the item parameters for a specific item can be found by solving a system of three equations in three unknowns, since item parameter triples are independent of all other triples.

---

[4] Refer to Hambleton and others (1978) for a review of several estimation techniques.

Lord (1980; Wood & Lord, 1976) suggests a modified Newton-Raphson procedure to perform the parameter estimation. It is an iterative procedure (using within each stage another iterative procedure) where a system of $N+3n$ nonlinear equations[5] is to be solved.

The iteration process is carried out in two stages. In the first stage the item parameters a, b, and c for all items are assigned starting values and estimates of ability (Θ) are computed for all examinees. This results in N linear equations being solved, one for each unknown (Θ). In stage 2 of the estimation procedure, the N Θ's are taken as fixed and the item parameters a, b, and c are estimated. At this point n sets of 3 equations in 3 unknowns (a, b,and c) are solved. These two stages are then repeated until convergence has been achieved.

This procedure is simplified greatly if the item parameters are already known as a result of pretesting[6]. In this case maximum likelihood estimation is used only to obtain the ability estimates of the examinees. This is done using an iterative process to solve for the critical value (Θ) of the likelihood equation for each examinee.

---

[5] N is the number of examinees required and n is the number of items.

[6] Recall from a previous discussion of ICCs in section 2.1.3.4, that ICCs are theoretically invariant across groups of examinees; hence the parameters defining the ICCs are also invariant.

A shortcoming of the theory behind the MLE procedure, at least in the context of this study, is the lack of a finite estimate of ability for response patterns consisting entirely of ones or zeros. A response vector of incorrect responses (all zeros) has an indeterminate ability level as does a vector of correct responses (all ones).

One solution to this problem is to employ a Bayesian method of ability estimation rather than a MLE method. This study employed an MLE procedure to estimate the item parameters and a Bayesian procedure to estimate ability levels (Θ).

Bayesian estimation procedures incorporate prior information about the distribution of ability in attempting to increase the precision of the estimation of Θ. With Bayesian procedures, it is generally assumed, a priori, that ability is distributed normally. This assumption results in the likelihood function being weighted by the standard normal density function. The weighted likelihood function is actually the joint distribution of Θ and item response vectors u. With the aid of Bayes theorem and some calculus, this distribution is transformed into the conditional distribution of Θ given a response vector u. This result is referred to as the posterior distribution of Θ given u.

The estimate of an individual's ability is simply the first moment (mean) of this distribution. Numerical methods must be employed in determining the estimate, as closed, computational forms have not yet been developed. A popular numerical algorithm to perform this task is Owen's (1975) Bayesian sequential procedure described in detail in section 3.3.2.

A subtle relationship exists between Bayesian estimation and maximum likelihood estimation. If one considers the prior distribution of Θ to be rectangular rather than normal over the range of ability from minus infinity through positive infinity, then Bayesian estimation yields results identical to those obtained through MLE. Stated another way, if the prior assumption of Bayesian estimation is omitted then the resulting procedure is nothing more than maximum likelihood estimation.

The use of prior information in an estimating procedure is a source of contention in the area of applied measurement. Bayesians feel that their brand of estimation is superior to MLE because more information about the population is incorporated into the process, with the result that the Bayesian estimates are more precise than those of an MLE procedure. Proponents of the MLE feel that these procedures are superior because they rest on weaker assumptions and they yield estimates that are population-free.

Samejima (1980) recently examined this controversy and was able to dispel the notion that Bayesian estimation was generally superior to MLE methods. She demonstrated the existence of biases resulting directly from inappropriate assumptions regarding the specific nature of the prior distribution. However, it was not concluded that MLE was generally better than the Bayesian methods nor that the Bayesian procedures were inappropriate for ability estimation. The findings of McKinley and Reckase (1981) generally support those of Samejima, but McKinley and Reckase extend their conclusions, perhaps prematurely, to suggest that MLE is superior to Bayesian methods in large scale tailored testing. Nothing conclusive in this regard can be found in the literature: hence the debate remains academic and a definite direction for further research.

2.1.7 Information

The concept of information is an important but seldom optimally used (Samejima, 1977) feature of latent trait theory. It is described by Bejar, Weiss, and Gialluca (1977, p. 2) as "an index of the precision of measurement at all levels of the trait being measured." Information relates the discriminating power of an item to the ability continuum, and for a given value of Θ, is inversely related[7] to the standard error of the estimate of Θ. Consequently, it follows that the smaller the variance at a particular ability level, the greater the information.

Information can be considered in the context of individual items (item information, $I(\Theta,u)$) or in the context of tests (test information, $I(\Theta)$). Item information is defined as a function of the item characteristic curve and its first derivative with respect to Θ. A computational definition of $I(\Theta,u)$ is given as follows:

---

[7] The relationship is actually $I(\Theta,u)=1/SE^2$.

$$I(\Theta,u)=[d^2(1-c)]/\{[c+\exp(d(\Theta-b)][1+\exp(-d(\Theta-b))]^2\} \qquad (4)$$

where:

a, b, and c are defined as in equation (2) and

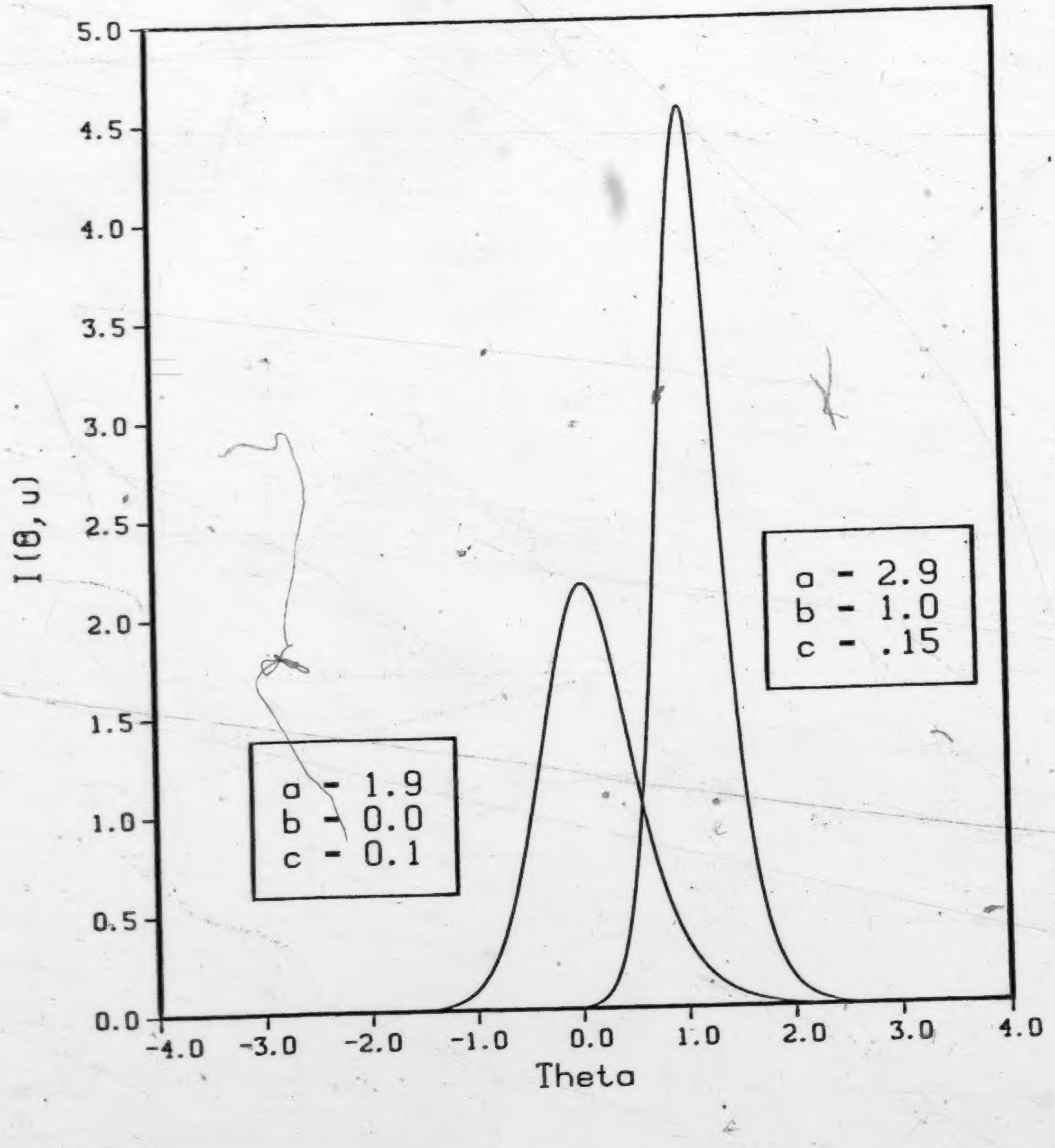
d is $1.7a$.

An item information curve is the graphical representation of the function $I(\Theta,u)$. It is defined for all values of Θ and is determinable for any item with known or estimated parameters a, b, and c. Examples of the information curves corresponding to the items in Figure 2.2 are found in Figure 2.5.

For the one and two parameter models as well as for the three parameter model when $c=0$, the item information curves $I(\Theta,u)$ are symmetrical about b. For the general, non-symmetric case of the three parameter model, the value of Θ that maximizes $I(\Theta,u)$ is always greater than (*i.e.* to the right of) b. The loss of symmetry is manifest by a positive skew which is caused by guessing at the lower values of Θ. This is so, as the effects of guessing are not uniform across the entire range of ability. Guessing is more prevelent among examinees at the lower end of Θ than among examinees at the upper end and seriously reduces the amount of information at that end of the ability range. It can be demonstrated that the higher the value of c the more skewed

Figure 2.5

Item Information Curves, I(Θ, u)



the curve becomes.

Lord (1980) considers test information $I(\Theta)$ to be a function of the independent and additive contributions of the items comprising the test. Birnbaum (1968) used the additive nature of the $I(\Theta,u)$ to define $I(\Theta)$ as the sum of the information curves corresponding to each item in the test:

$$I(\Theta)=\Sigma_j I(\Theta,u_j). \quad (5)$$

This function relates the discriminating power of the entire test to the ability continuum. The test information curve generated from a two item test[8] can be found in Figure 2.6. Lord (1980) has shown by way of a theorem that

> the test information function $I(\Theta)$ ... is an upper bound to the information that can be obtained by any method of scoring the test. (p. 71)

2.1.8 Applications

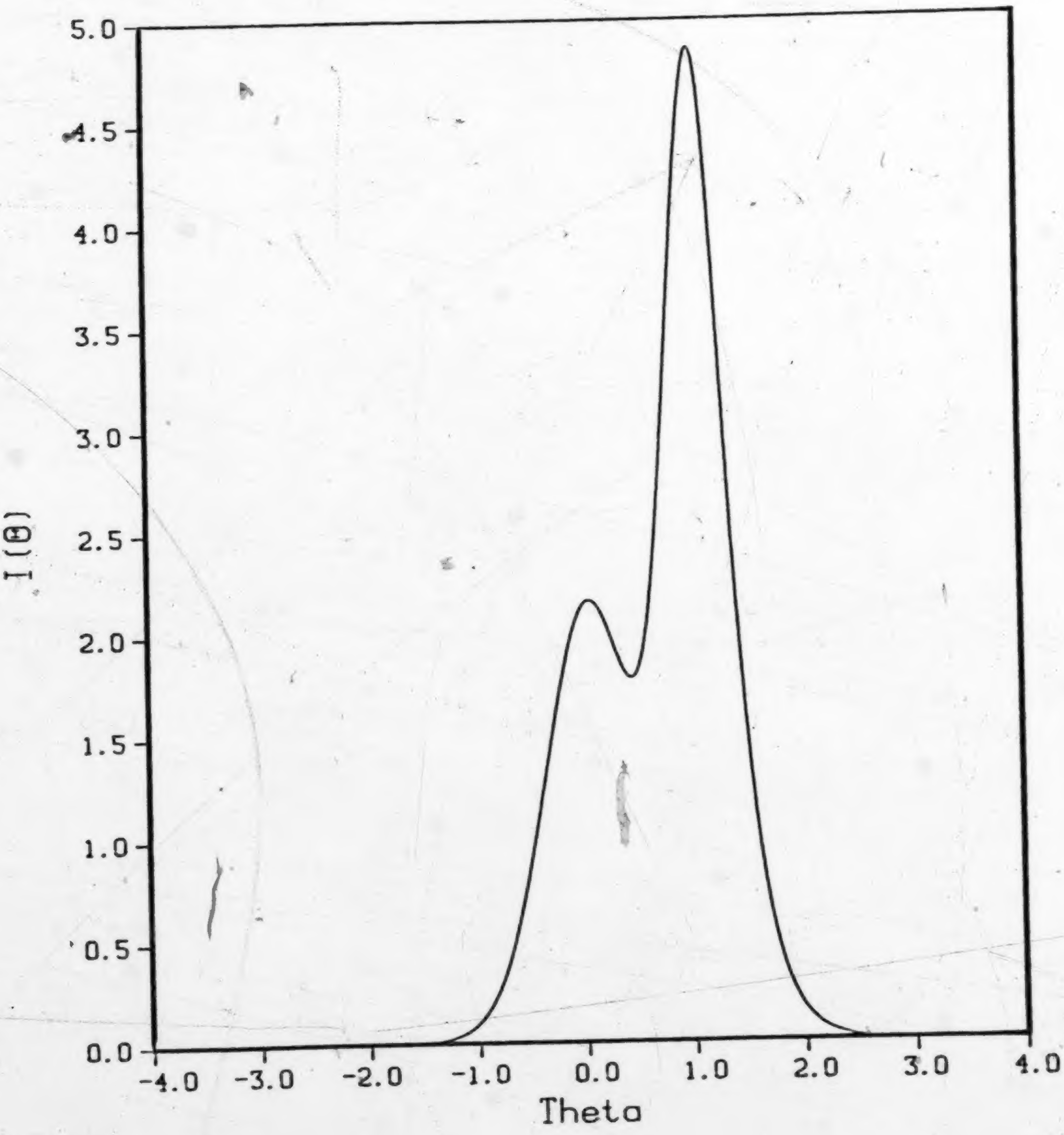
The number of areas to which IRT is being applied is increasing rapidly as its advantages become more widely known. Many applications are not new but rather attempts to improve upon the results obtained from other techniques. It is beyond the scope of this work to provide a detailed review of the above applications as such reviews can be found in Birnbaum (1968), Hambleton (1979), Hambleton, et al. (1978), Hambleton and Cook (1977), Lord (1977, 1980), Warm (1978), as well as other sources.

---

[8] Items from Figures 2.2 and 2.5.

Figure 2.6

Test Information Curve, I(Θ)



2.2 Tailored Testing

In an article entitled *Tailored Testing: A Successful Application of Latent Trait Theory*, Vern Urry (1977) made some very positive and exciting claims regarding the use of latent trait thoery. As suggested by the title of his article, he envisioned tailored testing as representing "a remarkably effective application of latent trait theory." (p. 181) He attributed tailored testing's success to the many major theoretical developments made in the area of test theory as well as the tremendous technological advances in computing (both in theory and power).

Recognizing the "impressive potential" (Urry, 1977, p. 181) of tailored testing Urry felt that a readable introduction to tailored testing was necessary to make it more widely known and practiced. His paper provided a general overview of tailored testing and contained suggestions for future development.

Urry presented several criteria that must be satisfied to derive maximum benefit from tailored testing. Among those listed, the choice of an appropriate latent trait model was considered to be of primary importance. Urry leaves no doubt that Birnbaum's three-parameter model is his preferred choice for tests consisting of multiple choice items. He justifies this by claiming that the effectiveness of tailored testing is decreased when models with one and two item parameters are used. Urry (1977, p. 184) dismisses the Rasch model as being "particularly inappropriate." He notes that even in the unlikely case where the assumptions underlying the Rasch can be met, the Birnbaum three-parameter model is still appropriate. The converse, however, is not true.

Kreitzberg, Stocking, and Swanson (1978) wrote a paper outlining some of the principles and theory basic to tailored testing, or "computerized adaptive testing" as they called it. In a discussion of the characteristics of conventional testing, Kreitzberg, et al. observed that most items, or at any rate a large proportion of them are expected to be of average difficulty for the target population. This in itself is considered a drawback to conventional testing since it results in measurements that are unequally precise for persons at different points along the ability scale. This result stems from the fact that measurement is more accurate and efficient when the items are approximately equal in difficulty to the ability of the examinees. Thus, in conventional testing where most items are of middle difficulty, the greatest precision tends to be at the center of the ability range, with precision decreasing toward both the extremes.

Hambleton (1979, p. 14) notes that in classical test theory, the variance of the errors of measurement is presumed to be the same across all ability levels. He suggests that this assumption is untenable and that what is needed are

> test models which can provide information about the

> precision of a test score (ability estimates) that is specific to and free to vary from one test score (ability estimate) to another. (p. 14)

Logic suggests that an obvious solution to this problem is simply to add easier and more difficult items to increase the number of items appropriate for individuals at the extremes of the ability range. However, this solution creates other problems such as causing testees at the upper end of the ability range to waste time on easy questions and confounding the individuals at the other end, causing them to guess or be frustrated or both by asking them to respond to difficult items.

Tailored testing provides a workable solution to this problem. The aim of tailored testing is to increase the precision of measurement of an individual by tailoring or adapting the test items such that the test consists of items that are appropriate for a given estimate of ability. This results in measurement of superior to or at least comparable with the precision provided by conventional tests at all points throughout the ability range, with the greatest improvement found toward the extremes.

In addition to asserting the superiority of tailored testing over conventional methods on the basis of precision and efficiency, Kreitzberg, et al. (1978) claim item economy as another major psychometric advantage. Their claim agrees with a statement by Lord (1970, pp. 173-174) and was substantiated by Urry, who found through experimental applications of a tailored test, an 80 percent reduction in the number of items required to achieve a level of precision comparable to conventional testing.

Further advantages listed by the authors refer more directly to the administrative and logistical aspects of computerized testing than the measurement properties of computerized tests. These include standardization of administrative procedures, reduction of administrator effects, and elimination of clerical errors in scoring.

Weiss and Betz (1973) published a review of tailored testing which summarized the research that had been done to that point in time. As Krietzberg et al. did in a later paper, Weiss and Betz partitioned the existing research into three methodological categories: empirical, simulative, and theoretical. The various approaches taken in tailoring test items to ability levels of individuals were considered within the constraints of the methodology employed. These approaches included two-stage procedures and a family of multi-stage strategies to which fixed and variable branching models belong.

The authors inferred that tailored testing possessed considerable potential for becoming the preferred method of the future and cited the following advantages:

1. considerably shorter than conventional tests, with little or no loss in validity or reliability,
2. more reliable than conventional tests in several

> studies and yielding more nearly constant precision than standard tests throughout the range of abilities, and
>
> 3. in several cases more valid, as measured against an external criterion. (p. 60)

Weiss and Betz considered Bayesian strategies, such as the one used in this study, as belonging to a class of variable branching models. The remaining discussion of tailored testing is directed toward this approach.

Weiss and Betz noted in their review that research with variable branching strategies had been limited, and reported a study by Novick (1969) describing one of the earliest applications of Bayesian logic. Novick (1969) proposed a model in which a weighted regression was incorporated to estimate an individual ability. The regression depended on an individual's observed score and the average observed score of a random sample of people from the population. During the initial stages of testing little information is known about an individual so it is assumed that he has the characteristics of a person selected randomly from the population, hence information on the population is weighted heavily. As the testing procedure progresses and more becomes known about the individual, less emphasis is placed on the population data and more weight is placed upon the information obtained from the individual (*i.e.* observed score).

Novick proposed that the item selection procedure be designed such that the item yielding the maximum amount of information, $I(\Theta,u)$, at the current ability level be chosen. This is in line with Lord's (1953b) comment and is fundamental to several item selection strategies used in subsequent research including the present study. To facilitate item selection, Novick suggested the use of a computer-based, interactive approach so that estimates of parameters can be continuously updated as each response is entered.

Using Monte Carlo procedures, Jensema (1974) examined the fidelity of Bayesian tailored testing by comparing two basic termination criteria:

1. Terminate testing when the standard error of estimate reaches a specific criterion level.
2. Terminate testing when a specific number of items have been administered.

The comparison involved the accuracy of the final estimate which was defined as the correlation between the final ability estimate and a pre-assigned "true ability" from which binary response vectors were generated. Moreover, the comparison was done separately for each of four imaginary item pools. Each pool consisted of 100 equally discriminating items. A different discrimination level was used for each pool.

Results showed that to terminate testing when a specific level of standard error had been reached was the better of the two methods under review. It provided a good index of fidelity independent of item discrimination, however as it was pointed out, the lower the item discrimination the larger the number of items required to achieve a given level of validity. It was determined that when a specific number of items was used as a termination criterion fidelity was highly dependent upon item characteristics. The authors demonstrated that the number of items required to achieve a specific level of fidelity could be estimated if the pool consisted of items of similar discriminating power.

Thompson and Weiss (1980) assessed the criterion-related validity of a Bayesian tailored testing strategy (Owen, 1975) against several criteria. This study was one of the first live-testing validation studies of tailored testing and was directed more toward practical application than theoretical development. In this study 131 college volunteers were administered a 40 item conventional, five alternative multiple choice vocabulary test. Each subject was also administered a comparable tailored test consisting of items selected from a pool of 200 items. (The actual size of the pool was 240 items, however items chosen for the conventional and tailored tests were to be mutually exclusive.) Dual termination criteria for the tailored test were adopted. The testing procedure was halted either when the standard error of estimate was less than or equal to 0.03 or when 135 items had been administered. The validation criteria included high school GPA, overall university GPA, university math GPA, and five scores from the American College Testing Program (4 subtests plus composite).

It was found that the correlations of the tailored test scores with the validation criteria were generally (in some cases significantly) higher than those relating the conventional scores to the same set of criteria. This would suggest that more valid ability estimates can be achieved from a Bayesian tailored testing strategy than from conventional testing methods.

The study uncovered a rather unexpected result. Contrary to the findings of research reported previously in this chapter, the tailored test was found to be 22 percent longer on average than the conventional test. Closer examination revealed that the median tailored test length was 12.5 percent shorter than the conventional test. This suggested that the majority of examinees responded to fewer items on the tailored test than on the conventional test, and that some examinees answered substantially more items on the tailored test. Thompson and Weiss attributed this finding to the large values of the initial prior variances' and a rather small termination criterion.

A slightly different approach to evaluating the merits of tailored testing was taken by Bejar, Weiss, and Gialluca (1977). They compared the information yield of a tailored test with that of a conventional classroom achievement test.

'Ranging from 2.0 to 4.0.

Data were collected by testing of over 700 students in a university level introductory biology course. Calculation of the corresponding information functions revealed substantially higher values of information across the range of Θ for the tailored test curve than the conventional test curve, hence more precise estimates of achievement were obtained via the adaptive test than the conventional one.

The foregoing information comparisons were based upon observed information functions. These curves were calculated as a function of the final ability estimates of the students who took the test. Given a student's final Θ, the item information functions (equation 4, p.42) were evaluated (at the particular ability estimate Θ) for each item administered to that student. These information values were then summed to obtain that student's "observed information value." This procedure was performed for each respondent and the resulting values were plotted to obtain observed information curves.

In the case of a conventional test, the foregoing procedure for calculating test information curves should yield observed information values corresponding to those for the theoretical[18] information curve. In the case of a tailored test, however, this procedure raises problems. First of all, test item information curves are, by definition, properties of tests and not people (*i.e.* calculated from item parameters rather than from an estimated Θ). Second, consider two subjects who have a common ability level but who have responded to entirely different sets of items. Two different sets of item information curves should be evaluated to obtain observed information values corresponding to the tests that these people took. Although the two ability estimates are equal, the information values may be different (since information is based on the item taken). In order to get an overall estimate of how well the tailored strategy works Weiss and colleagues averaged the test information curves obtained for all people at a specific Θ. Despite these problems, observed information curves were used in the present study because no better method for comparing tailored and conventional tests has been developed.

---

[18] These are calculated on the basis of the item parameters independent of ability estimates.

Recognizing that the conventional test might not have been psychometrically optimal, Bejar et al. (1977) constructed an "improved" version of the conventional test from the same item pool used to construct the previous instruments. The items for this test were selected in such a way that they were the most discriminating items in the pool. A comparison of the information functions for this and the tailored test produced results similar to those reported in the previous comparison, leading the authors to conclude that

> compared to conventional tests, adaptive achievement testing yields considerably more precise estimates of achievement, even when conventional tests are

> designed to take maximum advantage of the items in the pool. (p. 26)

The authors also examined the number of items administered during the testings. They found an average reduction of 30 percent through the tailored procedure, a far cry from Urry's claim of 80 percent but still impressive. The authors caution that a simple reduction of testing time may not necessarily be an advantage in itself unless ways of interfacing adaptive testing with the instructional process are developed that will increase instructional productivity.

Other studies (eg. Bejar & Weiss, 1978; Koch & Reckase, 1979; McKinley & Reckase, 1980) have generally supported previous research establishing tailored testing as an acceptable alternative to conventional testing methods. In satisfying a basic assumption of latent trait theory, the majority of studies have restricted themselves to relatively unidimensional item pools. However, in general, tests are not unidimensional.

Brown and Weiss (1977) addressed the problem of multidimensionality through a generalized tailored testing strategy for achievement test batteries. This strategy combined an inter-subtest branching procedure and an intra-subtest item selection strategy. Thus, multiple content areas could be tested without concern for the assumption of unidimensionality, provided the subtests were unidimensional. Item selection within each subtest was based upon the item information function. In particular the item chosen to be administered next was the item for which the information function was largest given the current estimate of ability, as determined by Owen's (1975) scoring procedure (section 3.3.2).

Branching between subtests was slightly more complicated in order to take advantage of available information about the interrelations among the subtests. The process involves two steps: subtest ordering and entry point calculations. Subtest ordering is determined by the raw-score inter-subtest correlations. The subtest chosen to be first is randomly selected from the pair of subtests having the largest zero-order correlation. The other subtest in this pair is chosen to be second. The remaining subtests are ordered on the basis of their multiple correlations with the subtests previously administered. Entry points to the second and subsequent subtests are calculated using a regression model for predicting ability on the nth subtest from the final ability estimates from the previous n-1 subtests. The entry point to the first subtest is arbitrarily made the same for all subjects.

Brown and Weiss evaluated their strategy by performing a real-data simulation on data collected from 365 fire control technicians at a naval guided missile school. The instrument consisted of 232 items, clustered into 10 subtests ranging in size from 10 to 32 items. All items were of the four option multiple choice variety. The evaluation was done in terms of three criteria:

1. Number of items administered for each subtest under each method.
2. Correlation between tailored ability estimates and conventional estimates for each subtest.
3. Observed test information curves, I(Θ), for each subtest.

Results from this study were favourable and similar to those obtained from previous studies which investigated tailored tests. The simulated tailored test required on the average 49.3 percent fewer items per subtest than did the conventional method. Pearson product moment correlations between the tailored test estimates of ability and the conventional estimates ranged between 0.74 and 0.98, with 11 of the 12 correlational values exceeding 0.90. The corresponding pairs of information curves were compared for each of the 12 subtests. Visual inspection of these curves did not reveal any appreciable difference between the pairs of tailored and conventional curves. The same general shape was apparent within each pair of subtest curves.

An empirical comparison of the pairs of the curves was performed. The ability range (*i.e.* the abscissa of the test information curve) was divided into intervals of 0.20 and the mean information value for each interval computed. Corresponding pairs of values were tested for significance by computing t-ratios. A few significant differences were found but a trend was not evident.

Brown and Weiss concluded that their tailored testing strategy for achievement test batteries was effective in reducing the number of items required with minimal loss of precision. This conclusion is not as exciting as some of those drawn from studies of unidimensional testing instruments, but the authors note that refinement of their inter-subtest branching procedure might produce improved results.

The adaptive strategy proposed by Brown and Weiss (1977) was re-evaluated in a study by Gialluca and Weiss (1979). They simulated the administration of a university level, multiple-choice general biology test using responses collected by conventional pencil and paper methods from 800 subjects. The test consisted of 100 items divided into five subtests of varying length. A similar set of criteria to that used by Brown and Weiss (1977) was used in the Gialluca and Weiss study.

Gialluca and Weiss (1979) made a slight modification to the inter-subtest branching strategy as outlined in the previous study. Rather than basing the inter-subtest correlations and regression equations upon the number-correct scores, the researchers used Bayesian ability estimates computed from the conventional test. Justification was not provided, but one can surmise that as Bayesian ability levels are to be estimated, it would be more natural to use Bayesian rather than conventional ability estimates in the computations needed to implement the tailoring procedure.

The results obtained by Gialluca and Weiss generally corroborated those reported by Brown and Weiss. High correlations between tailored test and conventional test scores were found and total test length was impressively reduced without noticeable loss of information.

Unlike Brown and Weiss, Gialluca and Weiss examined the effects of the inter-subtest and the intra-subtest strategies separately. They found that when each subtest was treated independently (*i.e.* only the intra-subtest strategy was used), subtest length was reduced by 16 to 30 percent. However, when the inter-subtest procedure was also used, test length was reduced by another one to five percent. It is unfortunate that Brown and Weiss did not provide a similar set of data for comparison.

Gialluca and Weiss concluded that these findings are supportive of the tailored testing strategy for test batteries. This strategy offers a means of reducing test length without decreasing the precision of measurement. More research is needed to extend these findings into other testing situations. The present study does this.

2.3 Canadian Cognitive Abilities Test

The Canadian Cognitive Abilities Test (CCAT) is a timed, group administered test that is designed to measure an "individual's ability to use and manipulate abstract and symbolic relationships" (Thorndike & Hagen, 1974). The authors' intent is the measurement of rational thought by emphasizing discovery of relationships and cognitive dexterity rather than by emphasizing knowledge of specialized or exotic content.

2.3.1 An Overview

Six levels of the test have been developed to measure the range of ability found in grades 3 through 9. Each level is comprised of three separate batteries - verbal, quantitative, and non-verbal - with each battery consisting of a series of subtests. The focus of the present study is limited to the verbal battery of the uppermost level (F), appropriate for grades 8 and 9. The verbal battery is made up of four subtests, which purport to measure competence in the following domains: vocabulary, sentence completion, verbal classification, and verbal analogies.

Each battery of the CCAT is presented in a multi-level format, with the items from all levels of each subtest being concatenated to form a single series beginning at level A and progressing to level F. The different levels are identified by varying the entry and exit points along the continuum of items.

2.3.2 The CCAT Test Package

The examiner's manual is well designed with particular attention given to the clarity of instructions for administration and interpretation. This should make the test a very useful tool to those unfamiliar with standardized testing. The pages of the multi-level test booklets are crowded with all the different entry and termination points marked. This could be a source of confusion to test takers.

A major deficiency of the CCAT test package is the lack of an accompanying technical manual. Correspondence with the CCAT's publisher, *Nelson Canada Ltd.*, revealed that a technical manual for the Canadian edition "is not yet available." The publishers did provide a copy of the technical manual for the U. S. edition (Cognitive Abilities Test), a series of handwritten tables (for the Canadian edition) and a claim that, with the exception of the tabular data, only minor differences exist between the two editions. The Canadian tables were used as the source of information for subsequent discussion of the CCAT's reliability, validity, and inter-subtest correlations. No information was available on such matters as practice effect and sex differences in performance.

2.3.3 Item Difficulty and Discrimination

Table 2.1 summarizes the estimates of discrimination power and difficulty for each item of the four subtests of the level F verbal battery. Difficulty indices are expressed as a proportion of correct responses while item discrimination is the biserial correlation between an item and the total subtest score.

SUBTEST	Item Difficulty		Item Discrimination	
	Median	Q	Median	Q
Vocabulary	0.57	0.23	0.53	0.09
Sentence Completion	0.77	0.19	0.64	0.07
Verbal Classification	0.70	0.09	0.53	0.06
Verbal Analogies	0.59	0.13	0.49	0.09

Table 2.1 Medians and Semi-inter Quantile Ranges of Item Difficulty and Item Discrimination

2.3.4 Norming

The CCAT was jointly normed with the Canadian Test of Basic Skills in the fall of 1973. A nation wide random sample of 189 schools was selected resulting in a sample size of approximately 26,700 students in grades 3 through 9. The sample was entirely English speaking and was stratified on the basis of school size, type of school, school organization, and province. Only national norms were established; regional norms were not developed.

2.3.5 Subtest Intercorrelations

From the normative sample the publishers drew 505 grade 8 and 9 (*i.e.* level F) test papers. Table 2.2 presents the resulting subtest intercorrelations of the verbal battery.

SUBTEST		A.	B.	C.	D.
Vocabulary	A.	1.00	--	--	--
Sentence Completion	B.	0.72	1.00	--	--
Verbal Classification	C.	0.67	0.66	1.00	--
Verbal Analogies	D.	0.66	0.67	0.62	1.00

Table 2.2 Inter-subtest Correlations

As noted previously, the CCAT's subtests are speeded, although they are intended to be power tests. The authors feel that because the items are graded by difficulty most examinees will have an opportunity to complete all the items they are capable of answering. This is not necessarily a correct assumption as speed will clearly affect the scores of slow subjects. The examiner's manual does not report any data regarding the effect of time on an individual's score.

2.3.6 Reliability

Data regarding the reliability of the CCAT are sparse. The only available estimate is a Kuder-Richardson Formula #20 (Kuder & Richardson, 1937) for the entire verbal battery: r_{11}=0.916. (The distribution of scores on which this coefficient is based had a mean of 63.11 and standard deviation of 13.89 raw score units.) A KR_{20} was calculated separately for each subtest and a formula for the correlation of sums was then applied to yield the resulting estimate for the entire test. Information as to the size and nature of the sample is unavailable. It should be noted that the KR_{20} is probably inflated due to the effects of speed on the test.

An estimate of the standard error of measurement was unavailable.

2.3.7 Validity

Criterion-related validity is the only aspect of validity for which information is provided. It consists of coefficients of correlation between the CCAT and the Canadian Test of Basic Skills. As noted in section 2.3.8, both tests were normed at the same time. Table 2.3 displays the correlations, which were based upon a sample of 496 papers drawn randomly from the norming sample.

CTBS SUBTEST	r
Vocabulary	0.79
Reading	0.79
Spelling	0.69
Capitalization	0.67
Punctuation	0.56
Usage	0.68
Language Total	0.76

Table 2.3 Correlations Between the CTBS Subtests and the CCAT Verbal Battery

2.3.8 In Summary

The lack of a technical manual for the CCAT severely limits any detailed, objective evaluation of the test. However, superficially, it appears that the CCAT may be useful to classroom teachers and other educators who find it necessary to make decisions about pupils.

The feasibility of tailored testing has been demonstrated for tests especially designed for that purpose, but it remains to be shown that comparable results can be obtained from imposition of a tailoring strategy on a conventional, standardized mental abilities test. The following chapter describes the methodology involved in assessing the effect of imposing such a strategy on the CCAT.

3. METHOD

The purpose of this study was to compare a conventional administration and a simulated administration of the verbal battery of the Canadian Cognitive Abilities Test. In brief, a set of item response vectors were dichotomously scored and then the items calibrated by the computer program LOGIST (Wood, Wingersky, & Lord, 1976). Using these parameters, conventional subtest scores (*i.e.* Owen's Bayesian Scores for the conventional administration) were calculated by the LINDSCO computer program (Bejar & Weiss, 1978) for each subtest, and tailored testing scores were calculated by the Simulated Tailored Tester program (SIMUTATER, see section 3.5.3.3, p. 82) again for each subtest. The two administrations were then compared in terms of estimated ability levels, number of items administered, and precision of measurement (*i.e.* test information). This chapter provides a detailed description of the methodology employed in the study as well as a restatement of the research questions of section 1.3.

3.1 The Instrument

The verbal battery of the Canadian Cognitive Abilities Test level F, was chosen for use in this study. The battery is comprised of the following four subtests: vocabulary, sentence completion, verbal classification, and verbal analogies. Each subtest is 25 items in length. A detailed discussion of the Canadian Cognitive Abilities Test (CCAT) can be found in section 2.3.

The test used in this study was selected after consideration of three criteria, two theoretical and one utilitarian in nature. The primary consideration was directly related to the general problem inasmuch as an already existing, standardized test was to be studied. The second criterion was that the test had to be suitable for analysis by a latent trait model. Specifically, the test had to yield dichotomous item scores, such as would be provided by typical multiple-choice tests. The final consideration was one of practicality. The Edmonton Public School Board was using the CCAT on a large scale and it was hoped that positive results from this study would demonstrate the potential of tailored testing to the Board.

The fact that the CCAT is slightly speeded necessitated a trade-off between practical use and model fit. It was decided that the need to demonstrate the practical applicability of tailored testing should outweigh concern over possible violations of the assumption of IRT, that the test is a pure power test.

3.2 Design

The paradigm followed in the study was that of a real data simulation. This type of study uses existing item level data gathered during an administration of the items as a conventional pencil and paper test. Relevant item statistics and parameters were calculated in preparation for the simulational process. An adaptive testing procedure was then simulated on the item data pool by imposing a tailoring strategy and rescoring each response set as if the testing had been carried out in an adaptive testing environment (Weiss and Betz, 1973).

Weiss and Betz (1973) point out that while such studies are obviously limited by the nature of the subjects and item pools, their strength lies in their control of subject by mode interaction -- all data are collected using the conventional testing mode. A disadvantage of real data simulations is that they cannot provide information on the actual effects of an adaptive testing strategy upon subjects.

Real data simulations should not be confused with the Monte Carlo study. Under the Monte Carlo paradigm hypothetical subject samples and item parameters are generated by computer under specified assumptions. Real data simulations, however, involve the rescoring of real examinee responses to real items, following a specific testing strategy.

3.3 Models

Two different models were employed in this study; a response model and a scoring or an ability estimating model. Each model is described in the following sections.

3.3.1 Response Model

As stated in section 2.1.5, the study was based upon the three-parameter Logistic Test Model (Birnbaum, 1968). For this model it is assumed that the probability of a correct response to a given item (*i.e.* $u=1$) increases as ability (θ) increases. In other words a monotonically increasing function is assumed. A further assumption is that this probability can be represented as a function of the three latent trait item parameters a, b, and c as follows;

$$P(1|\theta)=c+(1-c)[1+\exp(-1.7a(\theta-b))]^{-1} \qquad (6)$$

where:

a is the discrimination index,

b is the difficulty level, and

c is the pseudo-chance level for a given item.

All item calibration was performed using this model.

Related to this model, as discussed in detail in section 2.1.7, is the concept of item information, $I(\theta,u)$. The item information function is a function of θ and indexes the precision of measurement of an item across the ability range. It is defined once again in equation (7):

$$I(\theta,u)=[d^2(1-c)]/\{[c+\exp(d(\theta-b)][1+\exp(-d(\theta-b)]^2\} \qquad (7)$$

where:

a, b, and c are defined as in equation (6) and

d is $1.7a$.

An extension of this concept is the test information function, $I(\theta)$. It is defined as the sum of the information functions of the individual items that compose the test.

This function provides an estimate of the test's accuracy of measurement at a given Θ. All information analyses in this study were done using equation (7).

3.3.2 Scoring Model

Owen's (1975) Bayesian scoring model was used to estimate ability levels. This procedure is an application of Bayes' theorem to a stochastic process; the probable location of an individual's ability is re-estimated after each observation. The first central moment (*i.e.* expected value) of the derived (posterior) distribution of an individual's ability, given a response *u*, is considered to be the new, updated estimate of ability. Under this model, the new ability estimate is a function of the mean and variance of the prior distribution of an individual's ability.

Equations for the first and second central moments of the posterior distribution of Θ, given a normal prior distribution, $N \rightarrow (M_i, V_i)$, and given the observation (u_i), are as follows:

$$M_{i+1} = E(\Theta|1) = M_i + (1-c)V_i[\sqrt{(a^{-2}+V_i)}]^{-1}\phi(D)/A \tag{8}$$

$$M_{i+1} = E(\Theta|0) = M_i - V_i[\sqrt{(a^{-2}+V_i)}]^{-1}\phi(D)/\Phi(D) \tag{9}$$

$$V_{i+1} = V(\Theta|1) = V_i\{1-(1-c)(1+a^{-2}V_i^{-1})^{-1}\phi(D)[(1-c)\phi(D)/A-D]/A\} \tag{10}$$

$$V_{i+1} = V(\Theta|0) = V_i\{1-(1+a^{-2}V_i^{-1})^{-1}\phi(D)[\phi(D)/\Phi(D)+D]/\Phi(D)\} \tag{11}$$

where:

$\phi(D)$ is the normal density function,

$\Phi(D)$ is the cumulative normal distribution function,

$D=(b-M_i)[\sqrt{(a^{-2}+V_i)}]$,

$A=c+(1-c)\Phi(-D)$, and

a, b, and c are item parameters of section 3.3.1.

In this study an initial normal prior distribution with mean 0.0 and variance 1.0 (*i.e.* $N\rightarrow(0,1)$) was assumed. In other words an initial ability level of 0.0 was assumed for all subjects.

3.4 Population and Data Collection

The CCAT, level F is appropriate for students at the grade 8 or 9 level. Item response sets were collected through a grade and district wide testing program carried out by the Edmonton Public School Board in the fall of 1980. Response sets from a total of 4057 grade 9 students were collected for use in the study.

The test was administered by subtest so that all students took the same subtests in the same sequence. All subjects responded to four subtests, each consisting of 25 items. Data sets with missing items were eliminated as they were of no use to the study. This reduced the sample to 3453 students.

The population was then randomly split into two groups of 3000 and 453. The data from the larger group were used for item calibration and for computing the correlations and regression equations on which the inter-branching procedure was based. The data from the smaller group, the validation group, were used in the real-data simulation.

3.5 Procedure

Sections 3.5.1 through 3.5.3.3 provide a stepwise description of the procedures that were followed.

3.5.1 Item Calibration

Item calibration was carried out on data obtained from the calibration sample (N=3000) using the computer program LOGIST (Wood, Wingersky, & Lord, 1976). LOGIST was designed to estimate the item characteristic curve parameters of Birnbaum's (1968) three-parameter logistic model. For each item a discrimination index (*a*), a difficulty level (*b*), and a pseudo-chance parameter (*c*) were estimated. Maximum likelihood estimates for each individual ability were also calculated, but were subsequently discarded as they are irrelevent to this study. Since the items from each of the four subtests were to constitute four different item pools in the simulation, the items for each subtest were calibrated independently of the items of the other subtests. Hence it was necessary to repeat the calibration process four times, once for each subtest.

3.5.1.1 LOGIST

LOGIST employs the modified Newton-Raphson method as described in section 2.1.6 to perform parameter estimation. A discussion of the assignment of initial parameter values can be found on page 12 of the LOGIST manual (Wood, Wingersky, & Lord, 1976).

3.5.2 Rescoring Conventional Administration

In order to facilitate a comparison of the conventional administration and the simulated tailored testing administration the responses of the validation sample were rescored using Owen's (1975) Bayesian scoring procedure. Four ability estimates, one for each subtest, were obtained for the conventional administration of these subtests.

A similar rescoring of the response sets from the calibration sample was done and the results used to calculate the correlations and regression equations required for the inter-subtest branching strategy.

The scoring of the conventional tests was done using the computer program LINDSCO (Bejar & Weiss, 1979), which is an acronymn for LINear Dichotomous SCOring. This program was designed to score conventional tests in which examinees respond to all items. Dichotomous (0,1) response vectors are input and scored according to Owen's Bayesian scoring model as discussed in section 3.3.2.

3.5.3 Tailored Simulation

The simulative portion of the study provides for the imposition of a tailoring strategy upon the real data response sets gathered from the validation sample under the conventional administration. As noted previously, the strategy employed in this study consisted of both an inter-subtest and an intra-subtest branching component.

3.5.3.1 Intra-subtest Branching

This component of the simulation procedure determines which items are to be included in the tailored test result for each subtest, for each individual in the validation group. In a real tailored testing, this procedure would serve the function of item selection. Intra-subtest branching is the process which ultimately tailors or adapts the test to each individual's ability level. It is an iterative process that begins at the initial ability estimate used as a basis for item selection. The item selected is administered and the response then scored. On the basis of the response the ability level is re-estimated using Owen's (1975) procedure and another item is chosen. Iterations continue until one of two termination criteria is met. The initial ability estimate for the first subtest used in this study was set at a value of 0.0 for all examinees. As a result, the first item "administered" was the same for all examinees. The initial ability estimates for subsequent subtests are described in section 3.5.4.2.

Items were selected within subtests, such that at a given selection occasion, the item chosen is the one of all the items remaining in the pool that provides the maximum amount of information given the current estimate of ability. Once an item has been chosen, it is no longer available for selection and rescoring.

As noted in the previous paragraph, the concept of item information is crucially important to the process of item selection. The item information function, $I(\Theta,u)$, provides an indication of the accuracy of measurement for an item at a given ability level. It is therefore quite reasonable to want to administer items that afford the greatest precision of measurement. This is accomplished by selecting and administering the item with the largest information of all the items in the item pool.

The procedure of intra-subtest branching terminates when one of two conditions is satisfied. In the most obvious circumstance item selection is halted when all of the items in the subtest have been administered. In the second case, termination occurs when the information provided by each of the remaining items is less than an arbitrarily small, but predetermined value. The following sets of termination criteria were used in this study:

1. A minimum information level of 0.10 or exhaustion of the item pool.
2. A minimum information level of 0.05 or exhaustion of the item pool.
3. A minimum information level of 0.025 or exhaustion of the item pool.

4. A minimum information level of 0.01 or exhaustion of the item pool.
5. A minimum information level of 0.001 or exhaustion of the item pool.
6. No minimum information level but exhaustion of the item pool.

In all instances the "final ability" estimate of each subtest was the last ability estimate prior to termination.

3.5.3.2 Inter-subtest Branching

An inter-subtest branching strategy was employed to facilitate adaptive item selection between the four subtests of the CCAT verbal battery. The strategy used in this study was the Gialluca and Weiss (1979) modification of a proposal originated by Brown and Weiss (1977). This procedure is dependent upon the following two steps:

- subtest ordering, and
- differential subtest entry points.

To obtain maximum advantage from the subtest intercorrelations, the order in which subtests are administered is an important consideration. As considered by Gialluca and Weiss (1979) subtest ordering was determined by a linear regression of the four Bayesian ability estimates obtained from the conventional administration. A matrix of bivariate correlations was computed from the rescored calibration data, and the pair of subtests with the greatest correlation were then considered. One of these two subtests was randomly chosen to be administered first while the other was administered second. Next, multiple correlations based on the rescored calibration data were computed using each of the unadministered subtests as criterion variables and the first two subtests as predictors. The subtest having the largest multiple correlation with the predictors was administered third, while the remaining subtest was administered last. In the simulation procedure, every subtest was rescored in the same order for each response set.

Brown and Weiss (1977) termed "differential subtest entry points" to mean the initial ability estimate for entry into subtests other than the first. As noted in section 3.5.3.1, the initial ability estimate for (*i.e.* entry point to) the first subtest was assumed to be 0.0, hence all examinees were "administered" the same first item.

A differential subtest entry point is determined by the relationship between a respondent's estimated ability level and the intercorrelations (determined from calibration sample) among the subtests. They were calculated by estimating the regression equations on the calibration data and using the appropriate equation to predict the entry point to the next subtest from final ability estimates for the previously administered subtests.

Specifically, at the end of the first subtest, the bivariate regression equation (12) was evaluated to determine the initial ability estimate for subtest 2:

$$D_2=b_1\Theta_1+b_0. \qquad (12)$$

Equation (12) defines the entry point D_2 to subtest 2 from the weighted final ability estimate for subtest 1 (Θ_1). The variance of this estimate, which was required for the first ability level calculation during the intra-subtest branching procedure, was taken to be the squared standard error of estimate (equation 13).

$$SEE=\sigma_y\sqrt{(1-R^2)} \qquad (13)$$

The calculations of entry points into the subsequent subtests were merely step-like extensions or generalizations of equation 12. Regression equations (14) and (15) were evaluated to calculate entry points to the third and fourth subtests.

$$D_3=b_1\Theta_1+b_2\Theta_2+b_0 \qquad (14)$$

$$D_4=b_1\Theta_1+b_2\Theta_2+b_3\Theta_3+b_0 \qquad (15)$$

In the above equations, b_1, b_2, and b_3 represent multiple regression coefficients, with b_0 representing the regression constant. The initial prior variance for subtest 3 was the squared standard error of estimate for the multiple regression of subtests 1 and 2 on subtest 3, and similarily, the initial prior variance for subtest 4 was the squared standard error of estimate for the regression of subtests 1, 2, and 3.

3.5.3.3 Computer Simulation - SIMUTATER

A FORTRAN computer program, called SIMUTATER (acronym for SIMUlated TAilored TEsteR) was written for use in this study. It is a general program in the sense that a real-data simulation of an adaptive testing can be carried out for any set of test data.

SIMUTATER requires two data files, one containing item parameters for each item, and one containing the raw data response sets. SIMUTATER expects binary input data but can convert raw response data to binary form if the answer key is supplied. Regression coefficients must also be supplied.

The program yields one Bayesian ability estimate per subtest for each individual response set. Hence, a total of four scores were returned for each data set in this study. SIMUTATER includes an option which will output these scores to a printer, a disk file, or both. Also included is an option that allows the user to output a vector containing the code numbers of the items administered to each subject.

Both of the branching strategies discussed in sections 3.5.3.1 and 3.5.3.2 are incorporated in SIMUTATER. Figure 3.1 shows how these strategies are combined, and also indicates the general flow of the program. Following the rescoring of the first item, the intra-subtest branching strategy is employed until either of the two termination criteria is met. An entry

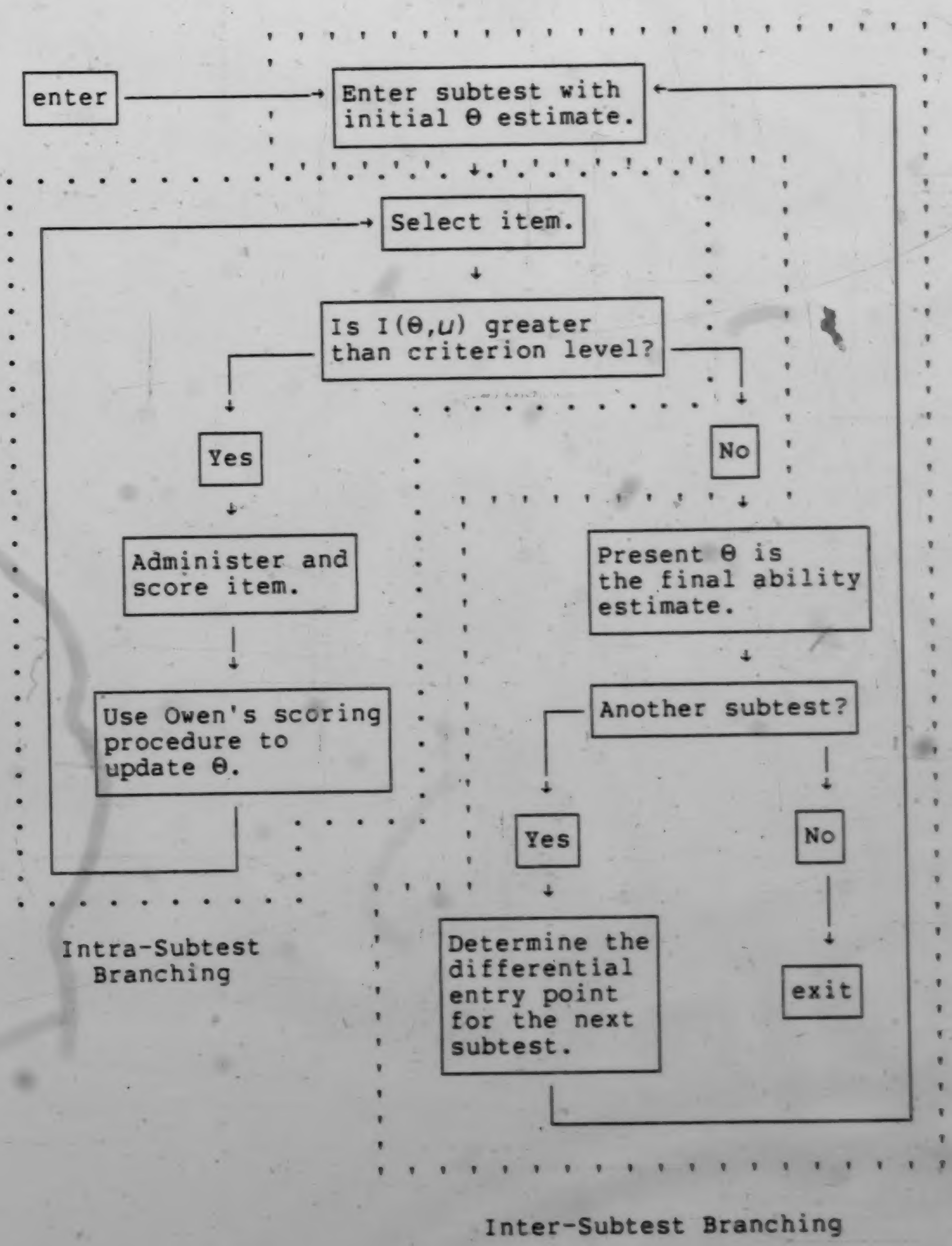

enter
Enter subtest with initial Θ estimate.
Select item.
Is I(Θ,u) greater than criterion level?
Yes
No
Administer and score item.
Use Owen's scoring procedure to update Θ.
Present Θ is the final ability estimate.
Another subtest?
Yes
No
Determine the differential entry point for the next subtest.
exit
Intra-Subtest Branching
Inter-Subtest Branching

Figure 3.1 Branching Strategies

point to the second subtest is then calculated using the inter-subtest branching strategy and the intra-subtest branching strategy is reapplied. The entire process is repeated until all subtests have been administered. At the termination of each subtest the final ability estimate is stored for later output.

3.6 Research Objectives

In this study the effect of imposing a tailor testing strategy on a conventionally administered test of mental abilities was examined under six different termination criteria and was assessed in terms of:

1. the correlation between IRT ability scores from the simulated tailored and conventional administrations of a subtest,
2. the number of items administered for a subtest when the tailored testing strategy was employed,
3. the correlation between IRT ability scores from the simulated tailored administration and the raw scores (*i.e.* number correct) for a subtest, and
4. the difference between the observed test information curves for the two administrations of the subtests.

The effectiveness of the inter-subtest branching strategy and the potential for increased precision at the extremes of the ability range were also examined.

4. Results

In attempting to answer the research questions posed in section 1.3 the study was conducted in the manner outlined in chapter 3. This chapter provides a detailed presentation of the results of the data analysis. It is divided into two sections: preliminary analysis, dealing mainly with the calculation of item parameters, and main analysis, reporting the results from the simulated tailored testing. Each portion of the results is discussed in turn.

4.1 Preliminary Analysis

This section is a report of the results from the procedures followed to obtain the item parameters and the other statistics necessary to perform the simulation. The response data analyzed in this section came from the calibration sample (N=3000).

4.1.1 Item Calibration

Raw item response data were used by the computer program LOGIST to compute the item response parameters for each item on each of the four subtests. As noted in section 3.5.1, the items in each of the subtests were calibrated independently of the items in the remaining subtests. The item parameters for each of the four subtests are presented along with summary statistics in Tables 4.1 through 4.4. They are again presented in Appendix A along with conventional difficulty and discrimination indicies.

Subtest A - Vocabulary			
Item	*a*	*b*	*c*
1	*0.598*	*-4.425*	*0.170*
2	0.664	-3.283	0.170
3	0.511	-1.098	0.170
4	*0.270*	*-5.531*	*0.170*
5	0.123	-0.738	0.170
6	*0.131*	*-9.005*	*0.170*
7	0.502	-2.930	0.170
8	*0.936*	*-3.775*	*0.170*
9	1.038	0.054	0.170
10	0.843	-2.127	0.170
11	*0.173*	*-6.904*	*0.170*
12	2.587	-1.092	0.170
13	0.752	-0.826	0.170
14	0.228	-0.197	0.170
15	0.259	2.621	0.170
16	0.987	-2.572	0.170
17	0.287	-2.528	0.170
18	0.883	-0.615	0.170
19	2.357	0.403	0.225
20	2.451	1.406	0.103
21	0.428	1.900	0.170
22	0.710	0.419	0.170
23	0.445	-0.493	0.170
24	2.717	0.783	0.140
25	2.442	1.639	0.200
μ	1.061	-0.464	0.169
σ	0.896	1.670	0.022

Note: Items shown in *italics* were removed from the study. The mean (μ) and standard deviation (σ) were calculated using only those items used in the simulation.

Table 4.1 Item Parameters for Subtest A

Subtest B - Sentence Completion

Item	*a*	*b*	*c*
1	*0.176*	*-13.746*	*0.130*
2	*0.476*	*-4.266*	*0.130*
3	0.454	-3.245	0.130
4	0.711	-1.704	0.130
5	0.668	-1.338	0.130
6	0.617	-3.129	0.130
7	*0.763*	*-3.623*	*0.130*
8	*0.840*	*-4.334*	*0.130*
9	0.685	-2.159	0.130
10	0.412	-2.638	0.130
11	0.883	-2.312	0.130
12	0.474	-2.078	0.130
13	0.668	-2.994	0.130
14	0.838	-1.737	0.130
15	1.118	-0.521	0.130
16	1.202	-1.643	0.130
17	0.775	-1.514	0.130
18	0.778	-1.094	0.130
19	0.630	-0.293	0.130
20	0.767	0.295	0.130
21	0.437	-3.158	0.130
22	0.403	-1.227	0.130
23	2.461	-0.006	0.150
24	2.565	-0.117	0.100
25	3.583	1.945	0.137
μ	1.006	-1.460	0.130
σ	0.830	1.330	0.008

Note: Items shown in *italics* were removed from the study. The mean (μ) and standard deviation (σ) were calculated using only those items used in the simulation.

Table 4.2 Item Parameters for Subtest B

Subtest C - Verbal Classification			
Item	*a*	*b*	*c*
1	*0.682*	*-4.152*	*0.185*
2	*0.244*	*-6.618*	*0.185*
3	0.357	-2.180	0.185
4	0.621	-3.024	0.185
5	0.423	0.357	0.185
6	0.665	-2.844	0.185
7	0.456	-2.882	0.185
8	1.136	-2.711	0.185
9	0.443	-1.811	0.185
10	2.841	0.064	0.192
11	1.232	-3.356	0.185
12	*0.406*	*-3.579*	*0.185*
13	2.701	-1.121	0.185
14	0.742	-1.025	0.185
15	0.685	-0.731	0.185
16	0.482	-1.049	0.185
17	0.338	-1.439	0.185
18	0.381	-0.433	0.185
19	2.758	1.122	0.114
20	0.632	1.966	0.170
21	0.633	-1.195	0.185
22	*0.092*	*-4.371*	*0.185*
23	0.394	-0.501	0.185
24	0.463	1.299	0.185
25	2.732	1.055	0.209
μ	1.005	-0.973	0.182
σ	0.995	1.553	0.017

Note: Items shown in *italics* were removed from the study. The mean (μ) and standard deviation (σ) were calculated using only those items used in the simulation.

Table 4.3 Item Parameters for Subtest C

Subtest D - Verbal Analogies			
Item	*a*	*b*	*c*
1	0.328	-1.843	0.145
2	0.915	-1.750	0.145
3	0.458	-2.523	0.145
4	0.746	-0.328	0.145
5	0.505	-0.978	0.145
6	0.547	-1.309	0.145
7	0.623	-2.820	0.145
8	0.417	-0.604	0.145
9	0.479	-1.506	0.145
10	0.419	-0.896	0.145
11	0.639	-3.301	0.145
12	*0.117*	*-7.199*	*0.145*
13	0.477	-1.600	0.145
14	0.411	0.540	0.145
15	0.784	0.538	0.145
16	0.802	-0.563	0.145
17	0.240	-2.174	0.145
18	0.708	0.540	0.145
19	0.322	0.465	0.145
20	2.477	-0.011	0.179
21	0.564	-1.471	0.145
22	2.256	-0.367	0.145
23	0.576	0.574	0.145
24	2.907	1.556	0.101
25	0.502	2.681	0.090
μ	0.796	-0.715	0.142
σ	0.702	1.423	0.016

Note: The item shown in *italics* was removed from the study. The mean (μ) and standard deviation (σ) were calculated using only those items used in the simulation.

Table 4.4 Item Parameters for Subtest D

An examination of Tables 4.1 through 4.4 reveals that the majority of *c* parameters were assigned a common value within each subtest. Lord (1974) points out that when the items are easy (*i.e.* low *b* values) the data do not provide for reasonable estimates of the *c* parameter as all subjects have a good chance of success. Lord continues that in order to "avoid absurd (although possibly harmless) estimates of *c*" (p. 258) with such data, LOGIST arbitrarily sets the *c* parameters equal to the median of the estimated *c* values according to a complex set of rules.

The information provided by an item is maximized if the examinee to whom it is administered has approximately a 50 percent chance of answering it correctly. As noted in section 2.1.7, the item should have a difficulty level (*i.e.* parameter *b*) slightly greater than the current estimate of the examinee's ability. Since ability estimates less than -3.50 or greater than 3.50 are rare it was decided that items in these extreme ranges of difficulty would be of little value, hence the subtest sizes were reduced to those shown in Table 4.5.

Jensema (1977) lists four characteristics that item pools should embody for efficient and precise tailored testing:

1. The items in the pools should be maximally discriminating ($a \geq 0.80$).
2. The items should have minimal chance (*c*) values.
3. The pools should be large.

SUBTEST		SIZE
Vocabulary (1)	A.	20
Sentence (2) Completion	B.	21
Verbal (4) Classification	C.	21
Verbal (3) Analogies	D.	24

Table 4.5 Final Subtest Sizes

[illegible]. The difficulty parameters of the items should be rectangularly distributed across Θ.

It can be seen from Tables 4.1 through 4.4 that the characteristics of the item pools studied here fall short of Jensema's criteria. This result, however, was not unexpected. The present study is addressed to the use of already existing item pools in tailored testing, not to the development of special pools that would be ideal in Jensema's sense.

4.1.2 Subtest Ordering

The subtest intercorrelations for the conventional IRT scores calculated on data from the calibration sample are presented in Table 4.6. By the procedure outlined in section 3.5.3.2, subtest A was selected to be "administered" first followed by subtest B.

SUBTEST		A.	B.	C.	D.
Vocabulary	A.	1.00	--	--	--
Sentence Completion	B.	0.62	1.00	--	--
Verbal Classification	C.	0.59	0.58	1.00	--
Verbal Analogies	D.	0.56	0.61	0.57	1.00

Table 4.6 Observed Inter-subtest Correlations

Multiple correlation coefficients for subtests A and B predicting each of the remaining two subtests C and D were examined. Subtest D was predicted marginally better (R=0.653) from subtests A and B than was C (R=0.651), hence subtest D was administered third and the remaining subtest, C, was administered last. The final order in which the subtests were administered was A, B, D, C. For the remainder of this report these will be referred to as subtests 1, 2, 3, 4.

4.1.3 Differential Entry Points

Section 3.5.3.2 explained that differential entry points to subtests would be calculated on the basis of regression equations for predicting the score of the next subtest from the final ability estimates of the previously administered subtest(s). The entry points to subtests 2 through 4 were computed by evaluating the appropriate regression equation from the set below.

$$D_2=.006+.597\theta_1; \qquad (16)$$

$$D_3=-.003+.295\theta_1+.427\theta_2; \qquad (17)$$

$$D_4=-.033+.275\theta_1+.248\theta_2+.253\theta_3. \qquad (18)$$

The variances of these estimates were taken to be the squared standard errors associated with the equations and are given below:

$$\sigma_2^2=0.432$$

$$\sigma_3^2=0.407$$

$$\sigma_4^2=0.367$$

As noted earlier the initial entry point into subtest 1 was zero (0.0) and with a variance of one (1.0) for all subjects.

4.2 Main Analysis

A comparison of the results obtained from the conventional test results derived through application of LINDSCO, with results obtained from the simulated tailored testing procedures derived through application of SIMUTATER, is presented in this section.

4.2.1 Termination Criteria

The simulated tailored testing procedures were repeated using six different termination criteria. A separate run of SIMUTATER was made for each criterion. The termination criteria used are listed below along with the symbol used to represent the ability estimates obtained under that particular termination criterion. These symbols are used throughout the remaining text.

Criterion		Symbol
0.100	–	C_{10}
0.050	–	C_{05}
0.025	–	C_{025}
0.010	–	C_{01}
0.001	–	C_{001}
0.000	–	C_{00}

The ability estimates obtained using the LINDSCO program will be referred to by "C" and raw scores will be symbolized as "R".

To get a feel for what the various criteria mean consider their relationship to the errors associated with ability estimates. If θ is estimated through maximum likelihood (MLE) procedures, then the standard error associated with estimate of a given θ is merely the reciprocal of the square root of the test information function evaluated at the particular θ (Hambleton, 1979). However, under Owen's Bayesian scoring algorithm the final posterior variance of the ability estimate is considered to be an estimate of the amount of error associated with that particular θ (Bejar & Weiss, 1979). As Bayesian scoring was used in this study the latter interpretation of error was used in the following discussion but the same principles would still hold if MLE procedures were used.

Urry (1977) writes that the objective of item selection strategies is to administer items in the order that will most rapidly increase information, or, equivalently, most rapidly decrease the error associated with the estimate.

Item selection is terminated when a criterion level is reached which specifies the minimum increase in information that an item may bring to the ability estimate. Alternatively the termination criterion is the minimum improvement (decrease) in error before item selection is terminated.

Higher levels of termination (eg. C_{10} or C_{05}) result in fewer items being administered, hence lower levels of measurement precision, larger posterior variances, and larger errors than those obtained when lower criterion levels are used (*i.e.* C_{01}, C_{001}, or C_{00}).

Practically speaking, little else is known about the nature or effects of the various criterion levels. Limited work using this kind of item selection strategy has been done by Weiss and his colleagues, with little attention paid to the specifics of the criterion levels. Typical criterion levels in their studies have been 0.05, 0.01, and 0.001.

4.2.2 Correlation Analysis

Pearson correlation coefficients were calculated for each subtest between the raw scores (*i.e.* number correct scores) and the Bayesian ability estimates obtained from the tailored testing simulations. The correlations found in Table 4.7 are for the ability estimates and the raw scores. At this point a matrix of low correlations would have raised serious questions concerning the appropriateness of the tailored testing strategy, but as all correlations in

Subtest	C_{10}	C_{05}	C_{025}	C_{01}	C_{001}	C_{00}
1	0.88	0.90	0.91	0.93	0.93	0.93
2	0.93	0.95	0.95	0.95	0.95	0.95
3	0.90	0.94	0.95	0.95	0.95	0.95
4	0.84	0.90	0.91	0.92	0.92	0.92

Table 4.7 - Correlations Between Raw Scores and Tailored Testing Estimates of Ability

Table 4.7 are high (22 of 24 greater than 0.90 and all greater than 0.84) the model was considered to be appropriate.

The Pearson correlations between the ability estimates from the conventional administration and those from the tailored testing for each termination criteria, are presented by subtest in Table 4.8. From this table several patterns become clear. First and most importantly all the correlations are extremely high. This indicated that the tailored testing procedure and the simple procedure of totalling the number of correct answers were yielding measurements on the same continuum. Second, the correlations between the tailored testing estimates of ability and the conventional estimates of ability increased as the termination criterion was relaxed (*i.e.* the criterion tended toward zero). Stated another way, as more items were administered in the tailored testing the correlation with the conventional score increased. This was not unexpected.

Subtest	C_{10}	C_{05}	C_{025}	C_{01}	C_{001}	C_{00}
1	0.96	0.98	0.99	0.99	0.99	0.99
2	0.97	0.98	0.98	0.98	0.99	0.99
3	0.96	0.98	0.98	0.99	0.99	0.99
4	0.92	0.95	0.96	0.97	0.98	0.99

Table 4.8 - Correlations Between Conventional and Tailored Testing Estimates of Ability

The highest correlations occurred when the termination criterion was set at 0.000 (C_{00}); that was when all items were administered under the tailored testing strategy. This situation differed from that of ability estimates obtained from the conventional scores in that the order in which the items were administered was almost certainly different from the order in which they were considered in the estimation of ability from the conventional administration. Local independence suggests that the correlations here should be 1.00, however, under Bayesian scoring the ability estimates are order dependent (Sympson, 1977). In other words a response vector scored by the Bayesian procedure and then re-scored after item rearrangement would produce two slightly different ability estimates. For this reason the correlations between C and C_{00} were less than 1.00.

Examination of Table 4.8 makes it clear that the imposition of a tailored testing strategy on the various subtests did not drastically change the estimation of an examinee's ability on a specific trait.

4.2.3 Comparison of Test Length

Table 4.9 displays the mean number of items administered per subtest under each tailored testing criterion. As the criterion became less stringent the average number of items increased, but in no case was the item pool exhausted on the average. For some examinees the item pools were exhausted. Reductions of subtest length ranged between 4% and 70% while reduction in total test length varied 8% to 58%.

Subtest	C	C_{10}	C_{05}	C_{025}	C_{01}	C_{001}
1	20	7.99	11.13	13.50	16.18	17.82
(σ)		(0.83)	(0.78)	(1.10)	(1.02)	(1.17)
2	21	11.21	16.63	18.55	19.45	19.87
(σ)		(3.34)	(3.94)	(2.43)	(1.28)	(0.66)
3	24	10.89	18.15	21.60	22.54	23.04
(σ)		(2.20)	(2.03)	(1.59)	(1.22)	(0.75)
4	21	6.20	14.36	16.94	17.76	18.72
(σ)		(1.26)	(1.84)	(1.15)	(0.94)	(0.89)
Total	86	36.30	60.27	70.62	75,95	79.47
(σ)		(6.01)	(6.67)	(4.67)	(3.19)	(2.45)

Table 4.9 - Mean and Standard Deviations of the Number of Items Administered

4.2.4 Efficiency Analysis

The theoretical test information curves are featured in Figure 4.1. This set of curves was generated by evaluating equation 7 (p. 44) for all items in each subtest. These curves indicate at which values of Θ the subtests have their greatest power of discrimination. The general profiles of these four curves reflect the strengths, and also the shortcomings of the item pools and were to some extent predictable through visual inspection of the distribution of the *b* parameters listed in Tables 4.1 though 4.4.

Figure 4.2 presents the averaged test information curves for the four conventionally administered subtests. As discussed in section 2.2 (p. 55-57) these are not true test information curves as they are calculated from observed Θ's as opposed to item parameters alone. They are estimates of the curves found in Figure 4.1. To form these curves (*i.e.* Figure 4.2) the ability scale Θ was divided into distinct intervals, 0.10 in width, and the informational values of examinees with ability estimates falling into an interval were averaged (see p. 54-55). This average was obtained for each interval. The resulting averages were then plotted against the mid-point of the corresponding Θ intervals. As expected the curves in Figure 4.2 closely matched the theoretical curves found in Figure 4.1.

Figures 4.3 through 4.6, one figure per subtest, present the averaged test information curves obtained under the simulated tailored testings. Each figure contains six

Figure 4.1
Theoretical
Test Information Curves

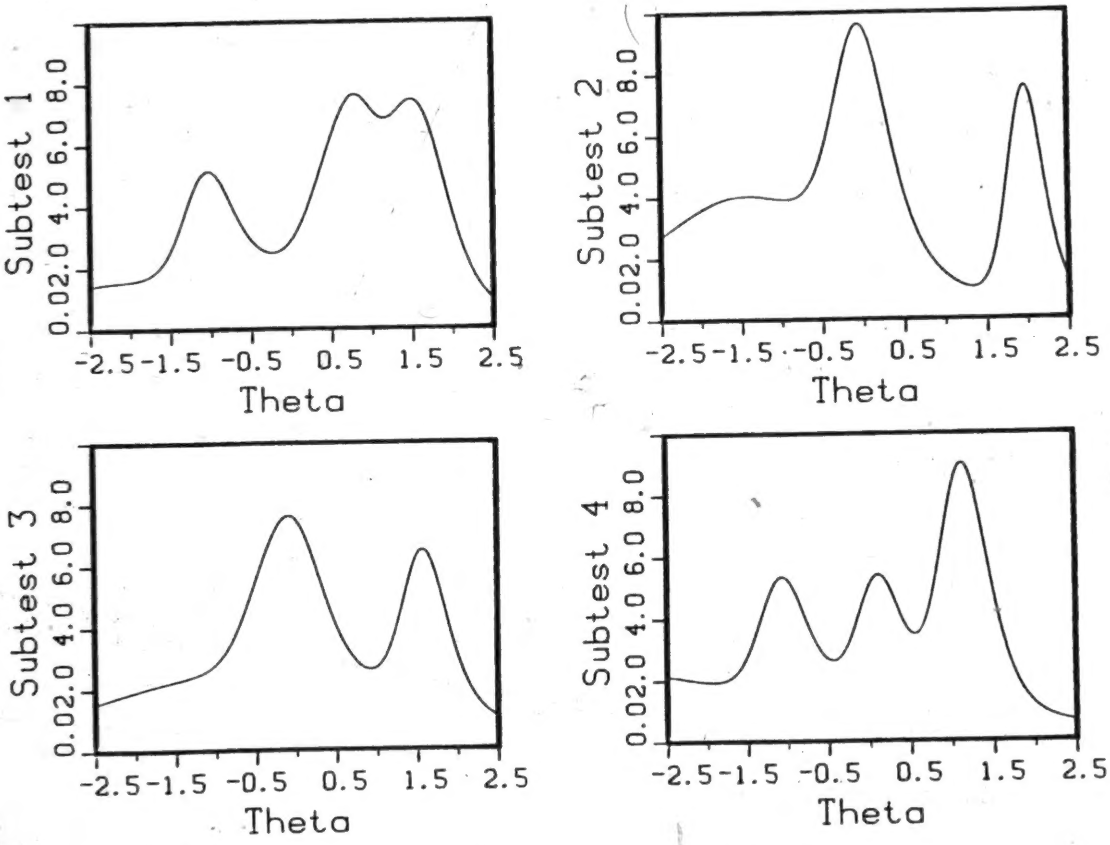


Figure 4.2
Averaged Conventional
Test Information Curves

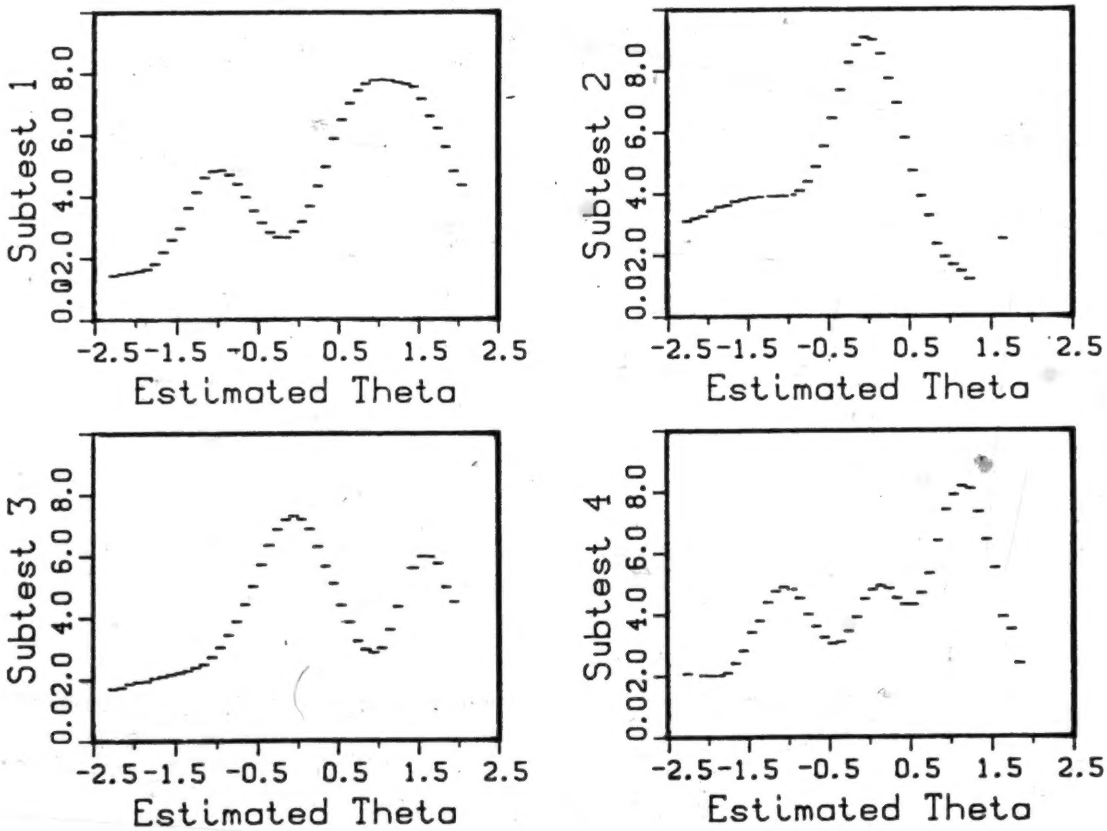


Figure 4.3
Averaged Test Information
Curves For Subtest 1

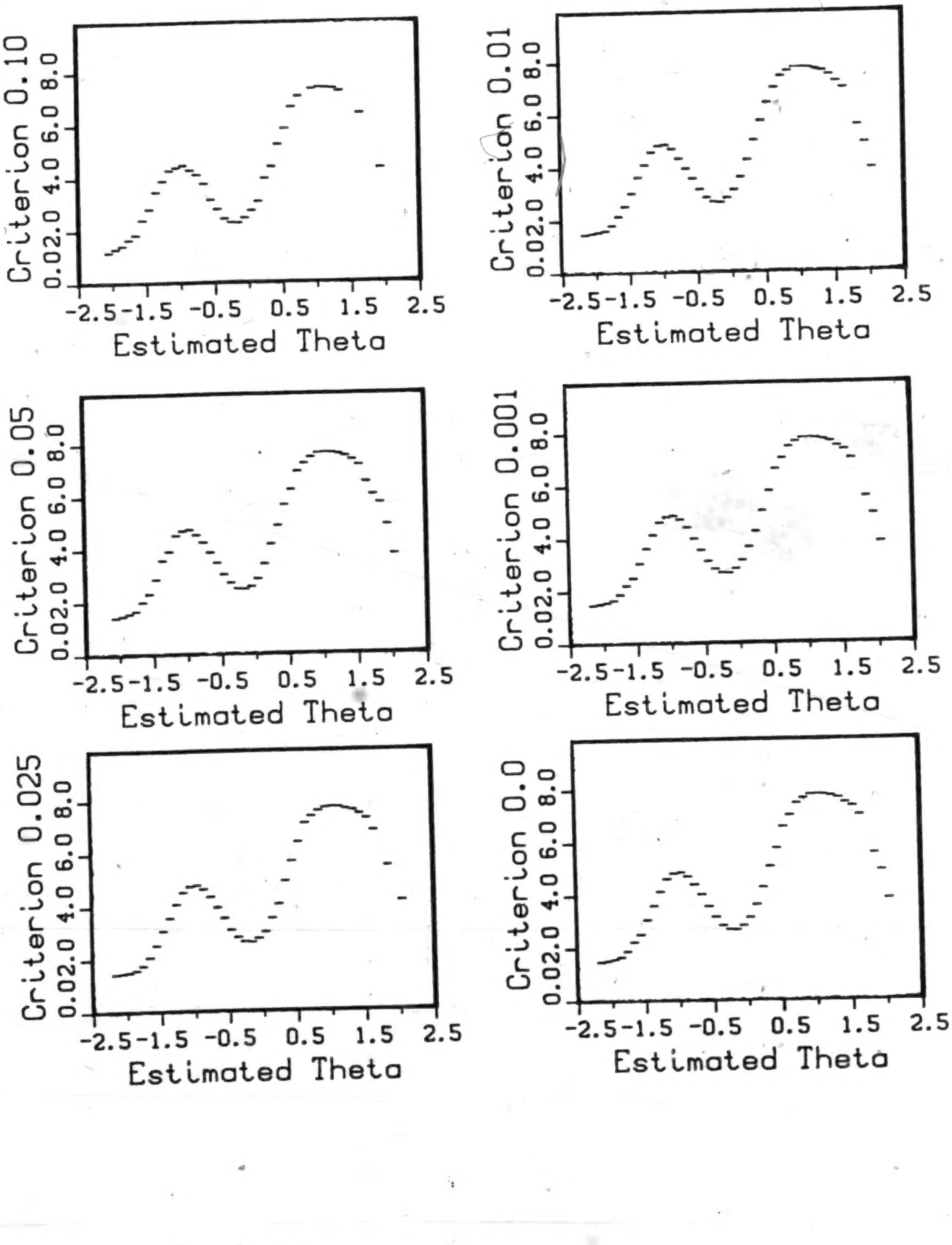


Figure 4.4
Averaged Test Information
Curves For Subtest 2

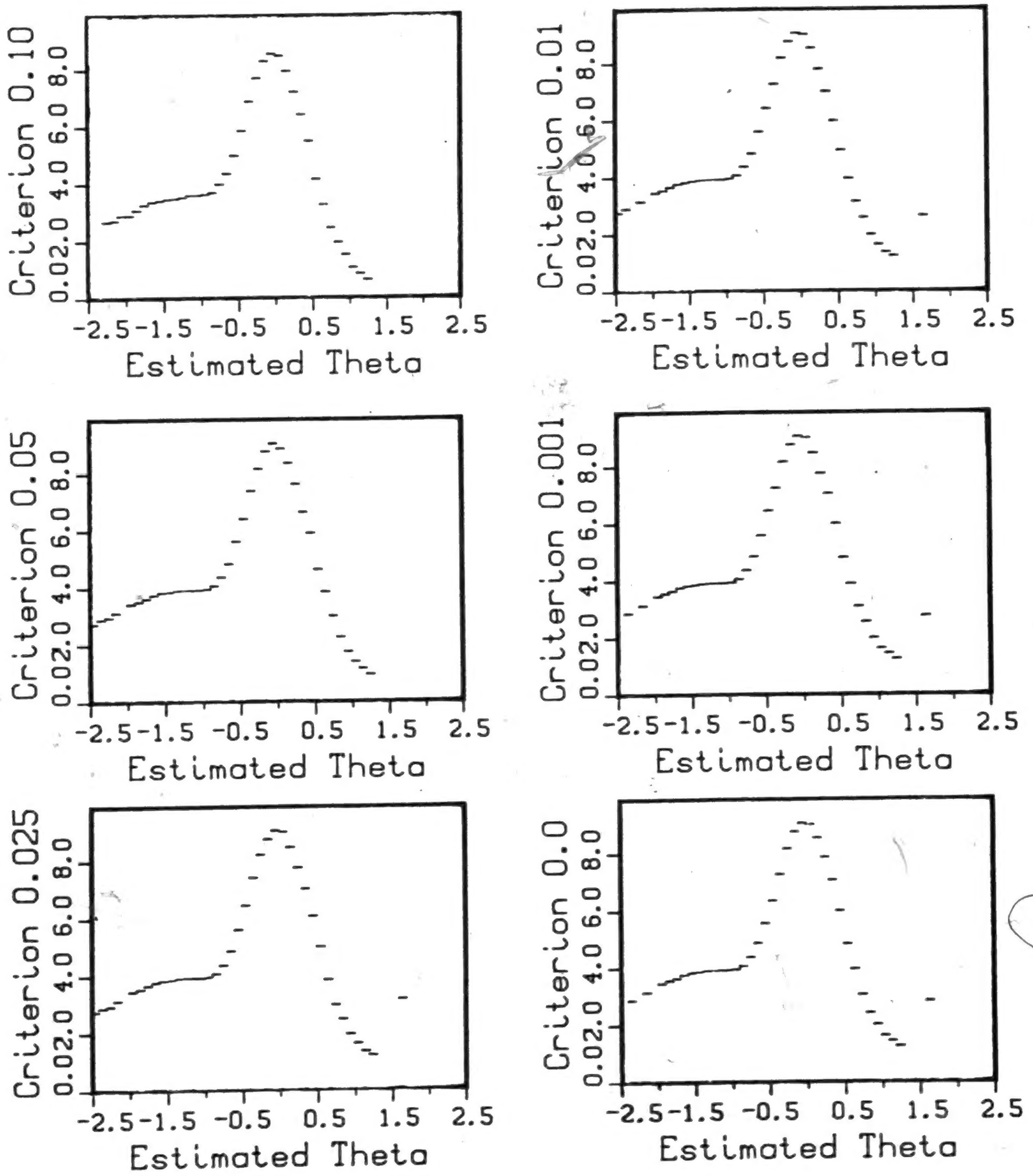


Figure 4.5
Averaged Test Information Curves For Subtest 3

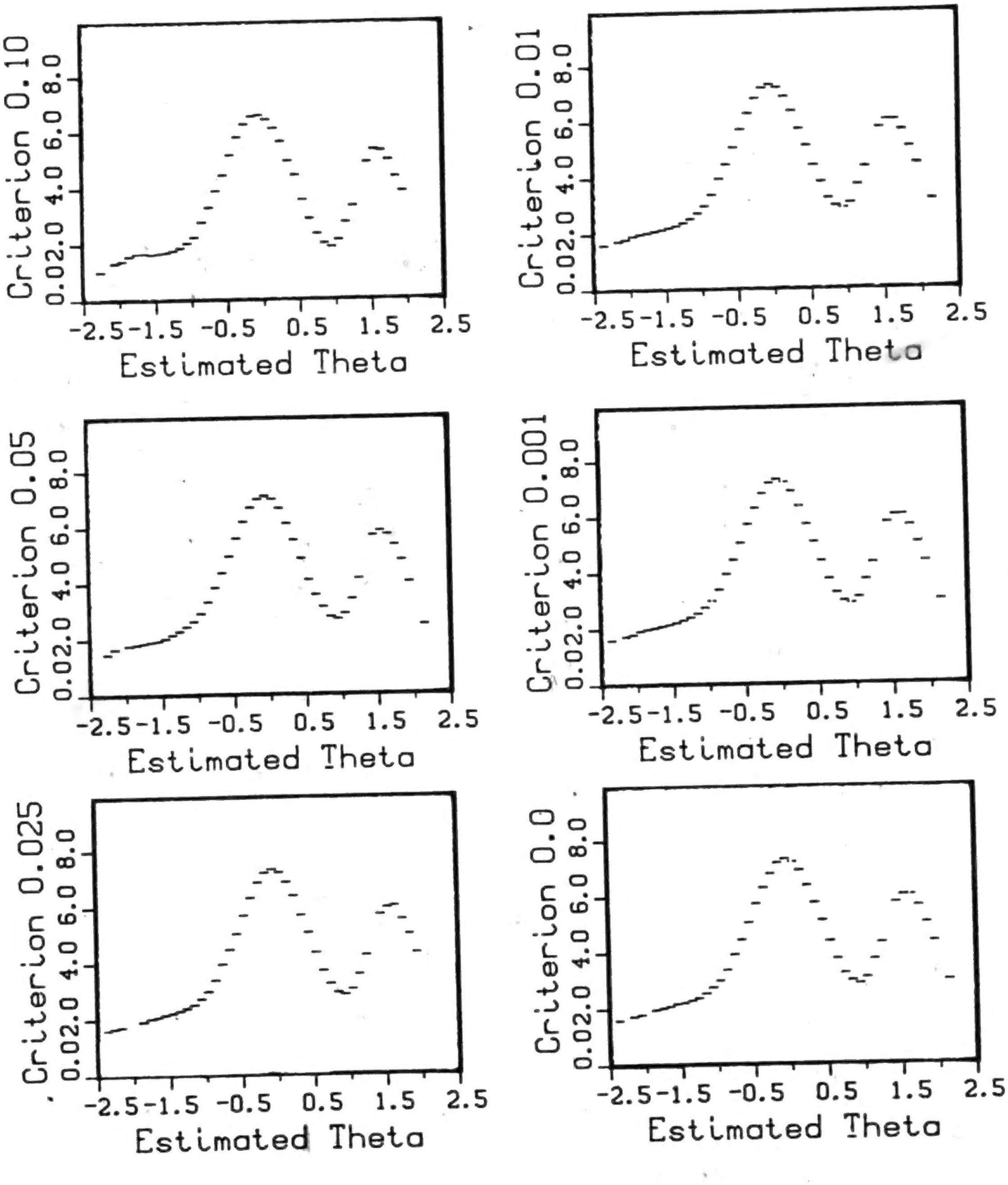
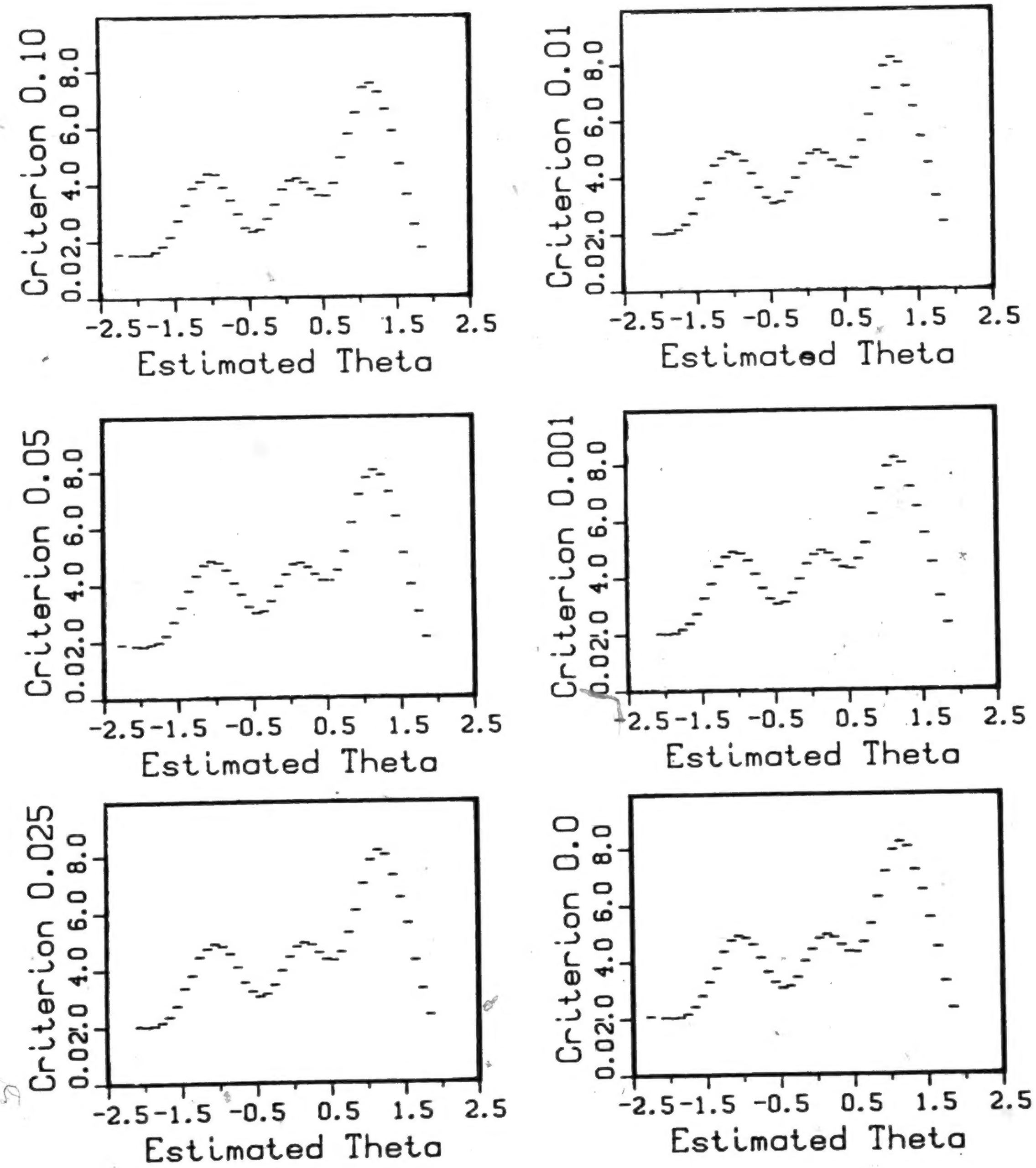


Figure 4.6
Averaged Test Information
Curves For Subtest 4



plots, one for each of the six termination criteria.

Considering each subtest by itself, the profiles of the curves are quite similar across the termination criteria. However, as the termination criteria became less stringent the curves moved slightly upwards. This indicated an increase in the amount of information as the termination criterion was relaxed.

This previous result was expected and explainable when considering the implications of relaxing the termination criteria. The termination criterion is the value at which testing will be stopped when there are no unadministered items remaining in the item pool that would provide an amount of information (at the current estimate of an examinee's ability, Θ) greater than the criterion. In other words, the termination criterion is the minimum amount of information that an item must have (at the current estimate of the examinees ability, Θ) in order to be administered. As the termination criterion becomes less stringent the number of items administered increases[11] hence increasing the amount of information obtained about the examinee's ability. (Recall information is additive.) Taken across the range of Θ, the entire test information curve rises.

A visual comparison of the general profiles found in Figures 4.3 through 4.6 with the appropriate curve in Figure 4.2 (and for that matter Figure 4.1) provides a quick way of assessing the effects of tailored testing. Throughout the series of figures the same four general profiles were found thus indicating that the subtests provided much the same pattern of precision under conventional and tailored strategies.

---

[11] Evidence for this statement can be found in Table 4.9.

Figures 4.7 through 4.10 display averaged posterior variances for intervals of Θ[12]. Comparing these figures to Figures 4.3 to 4.6 one important feature can be noted. In most cases intervals of Θ that possessed high information levels had averaged posterior variances that were at local minimums. This makes sense as high levels of information correspond to low levels of error in estimation. It can also be seen that as the termination criterion was relaxed, the magnitude of the posterior variances declined marginally.

Figures 4.11 through 4.14 are plots of the average number of items per Θ interval against the midpoint of the Θ interval. The data from which these curves were plotted are reported in Appendix B. Note that for the higher criterion levels (*i.e.* C_{10} and C_{05}) the shape of the curves are suggestive of their corresponding averaged test information curves. This is especially true for subtest 2. As the criterion level is relaxed the curve tends more toward a horizontal line, indicating less variablilty over Θ in the average number of items administered. Note that the intervals of Θ that had relatively large (relatively small) values of information also tended to have correspondingly small (large) average numbers of items administered.

---

[12] The data used to plot these curves are found in Appendix D.

Figure 4.7
Mean Posterior Variance vs Estimated Theta - Subtest 1

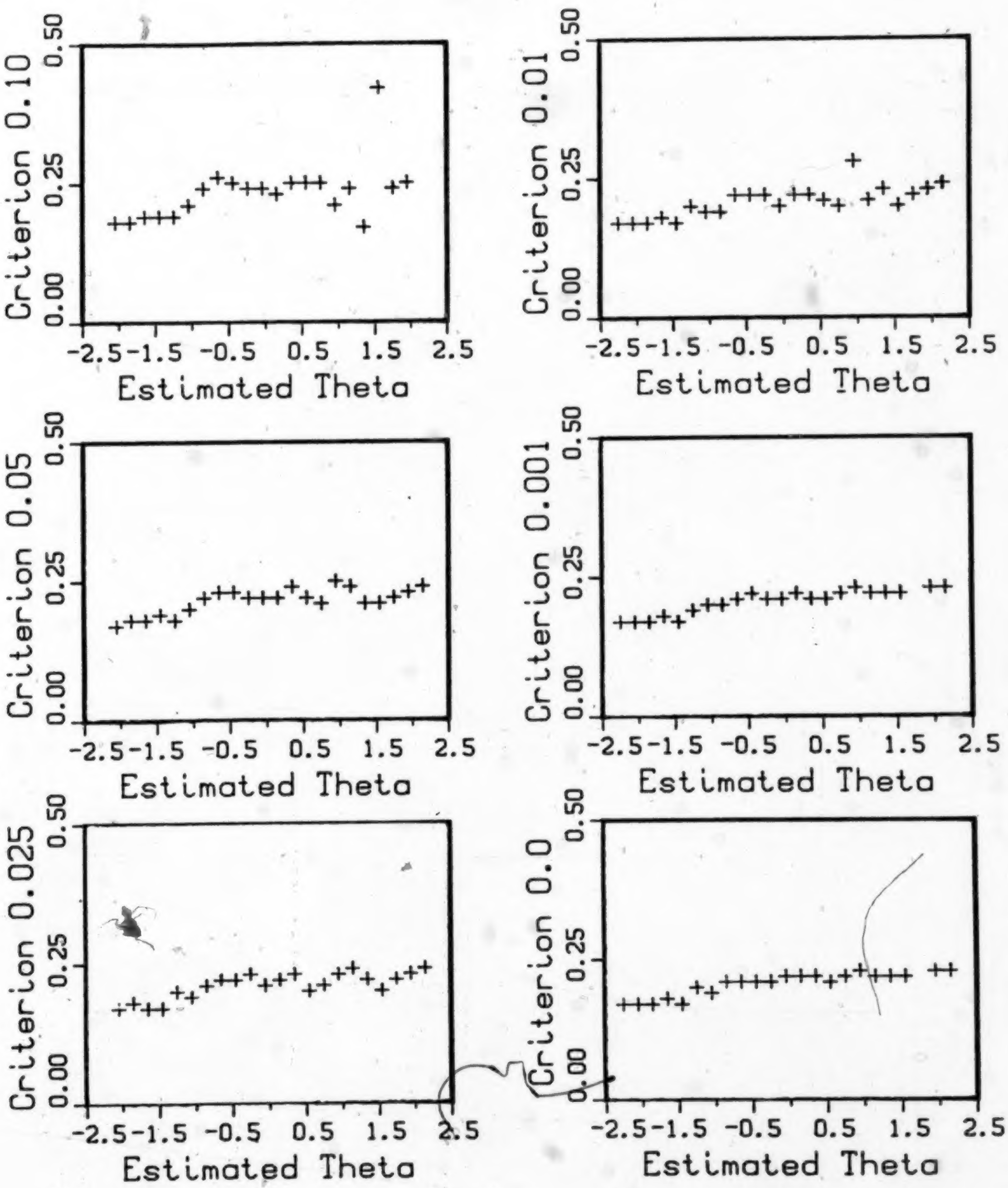


Figure 4.8
Mean Posterior Variance
vs Estimated Theta - Subtest 2

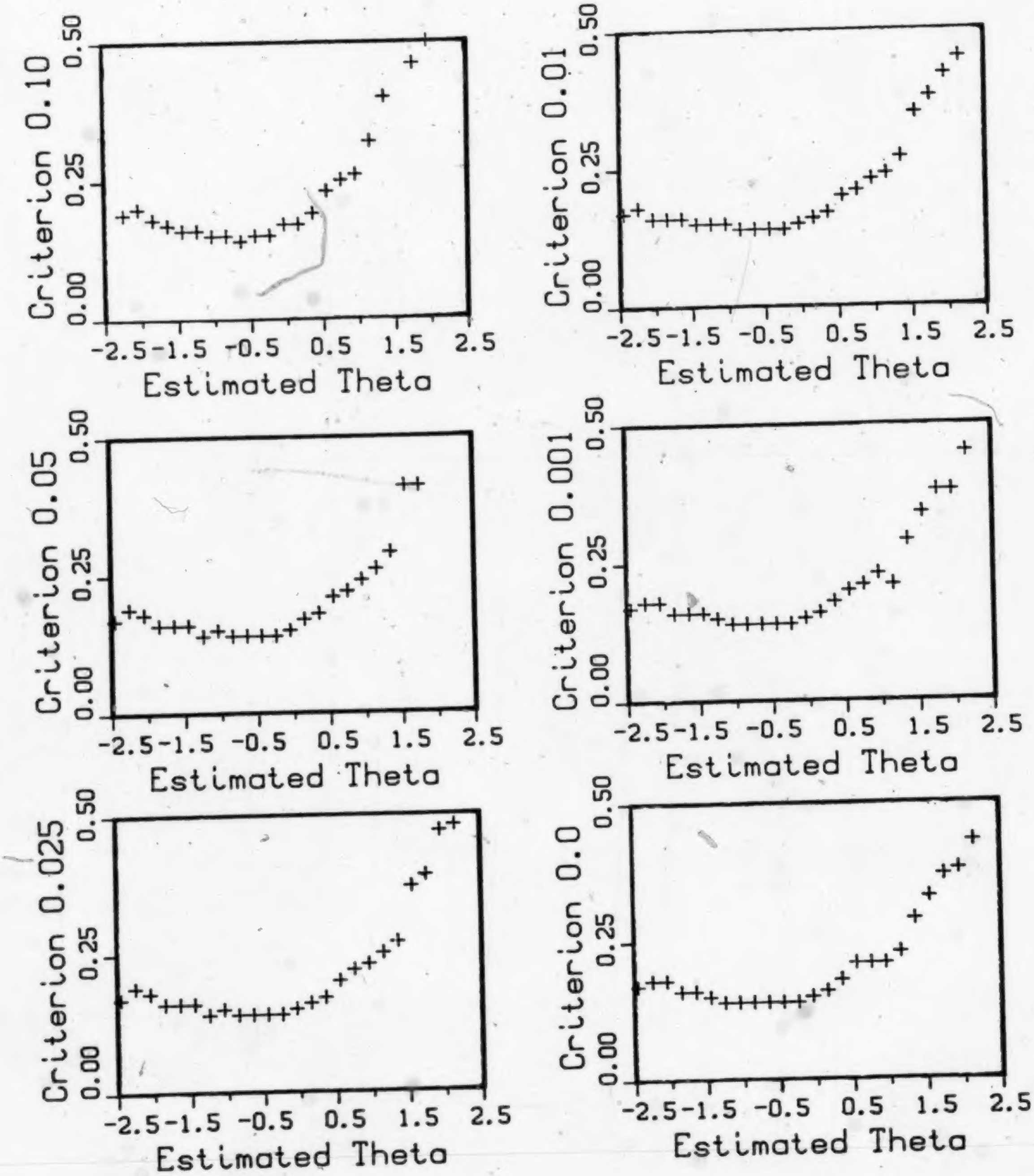


Figure 4.9
Mean Posterior Variance
vs Estimated Theta - Subtest 3

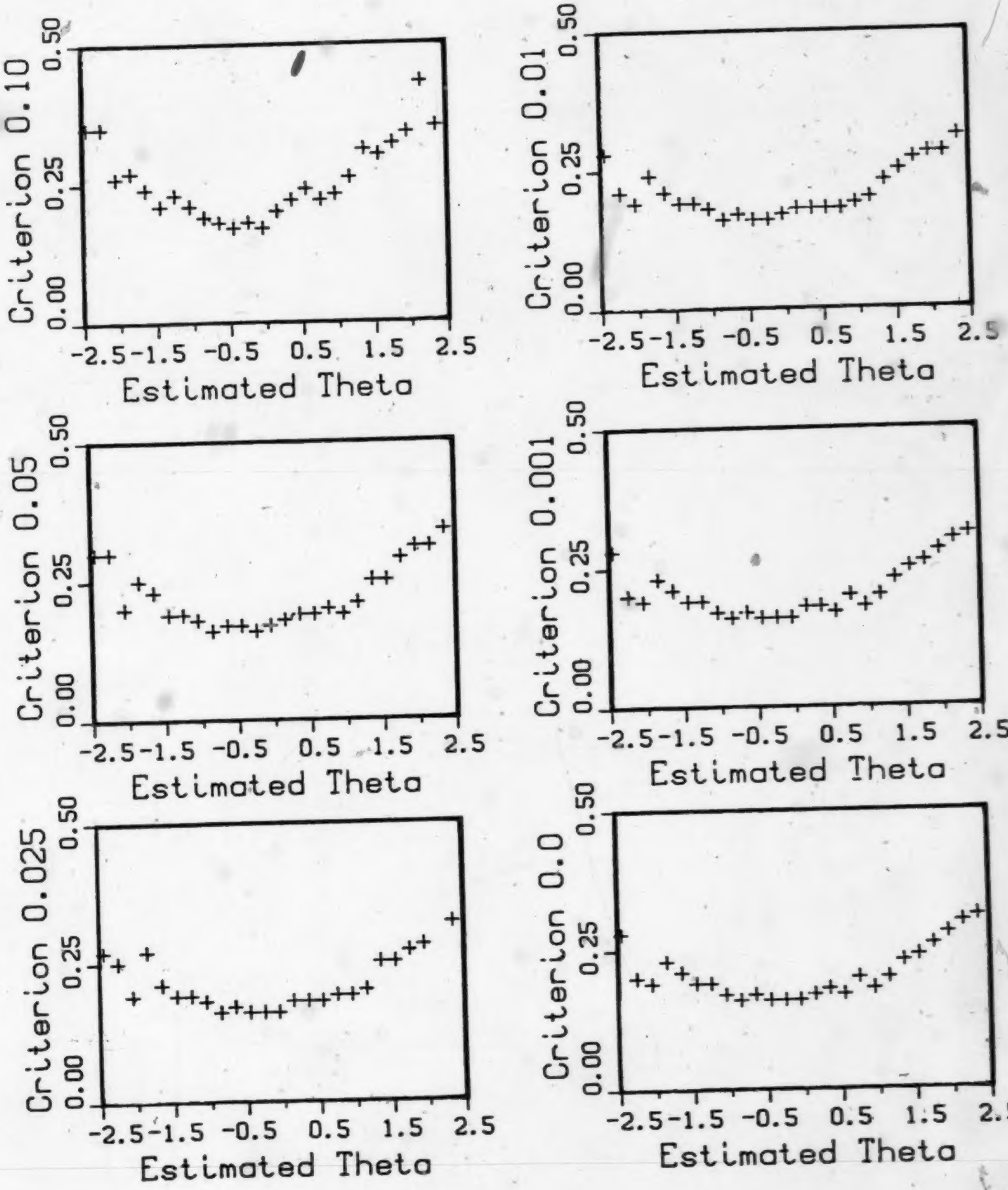


Figure 4.10
Mean Posterior Variance
vs Estimated Theta - Subtest 4

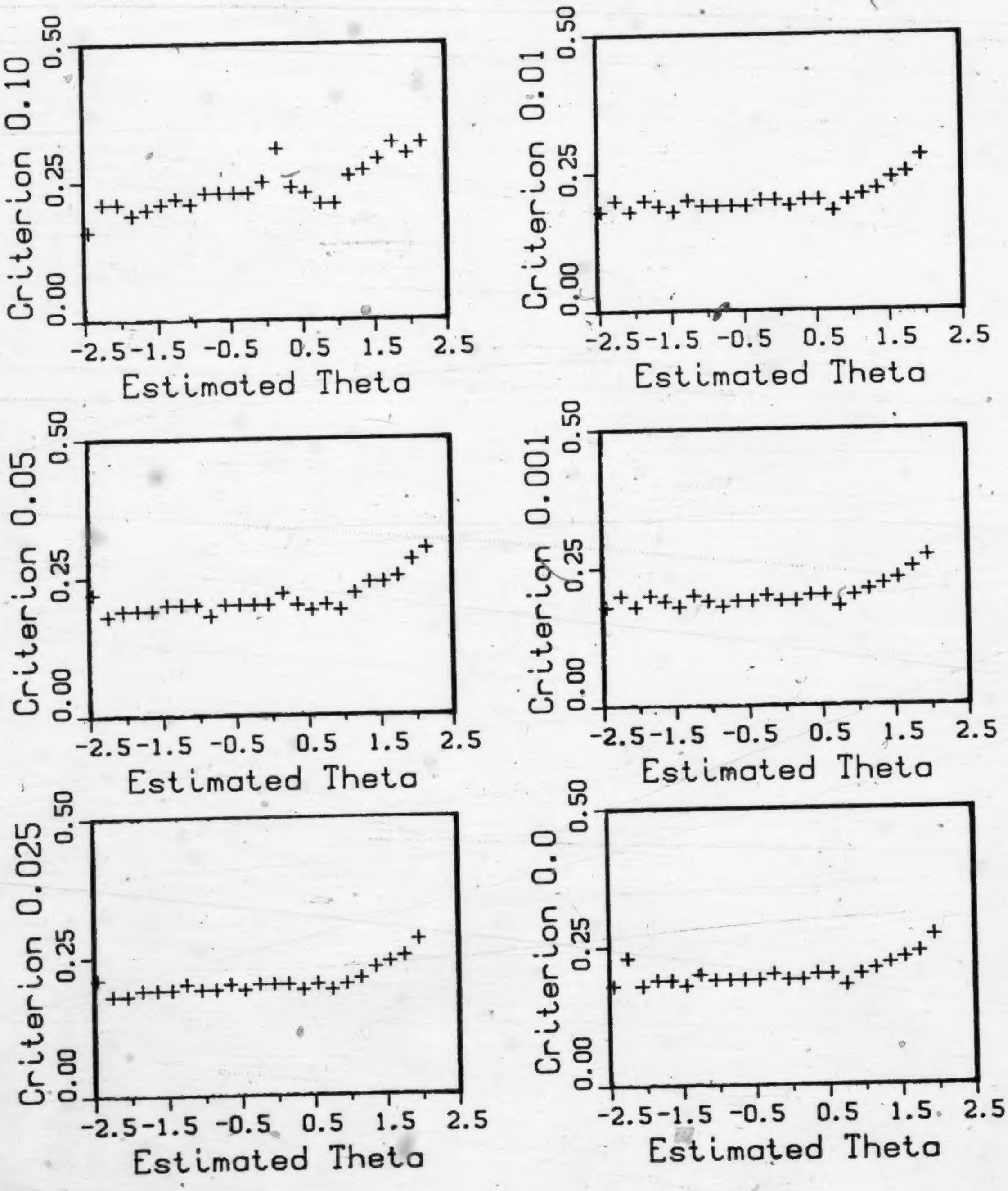


Figure 4.11
Mean Number of Items
vs Estimated Theta - Subtest 1

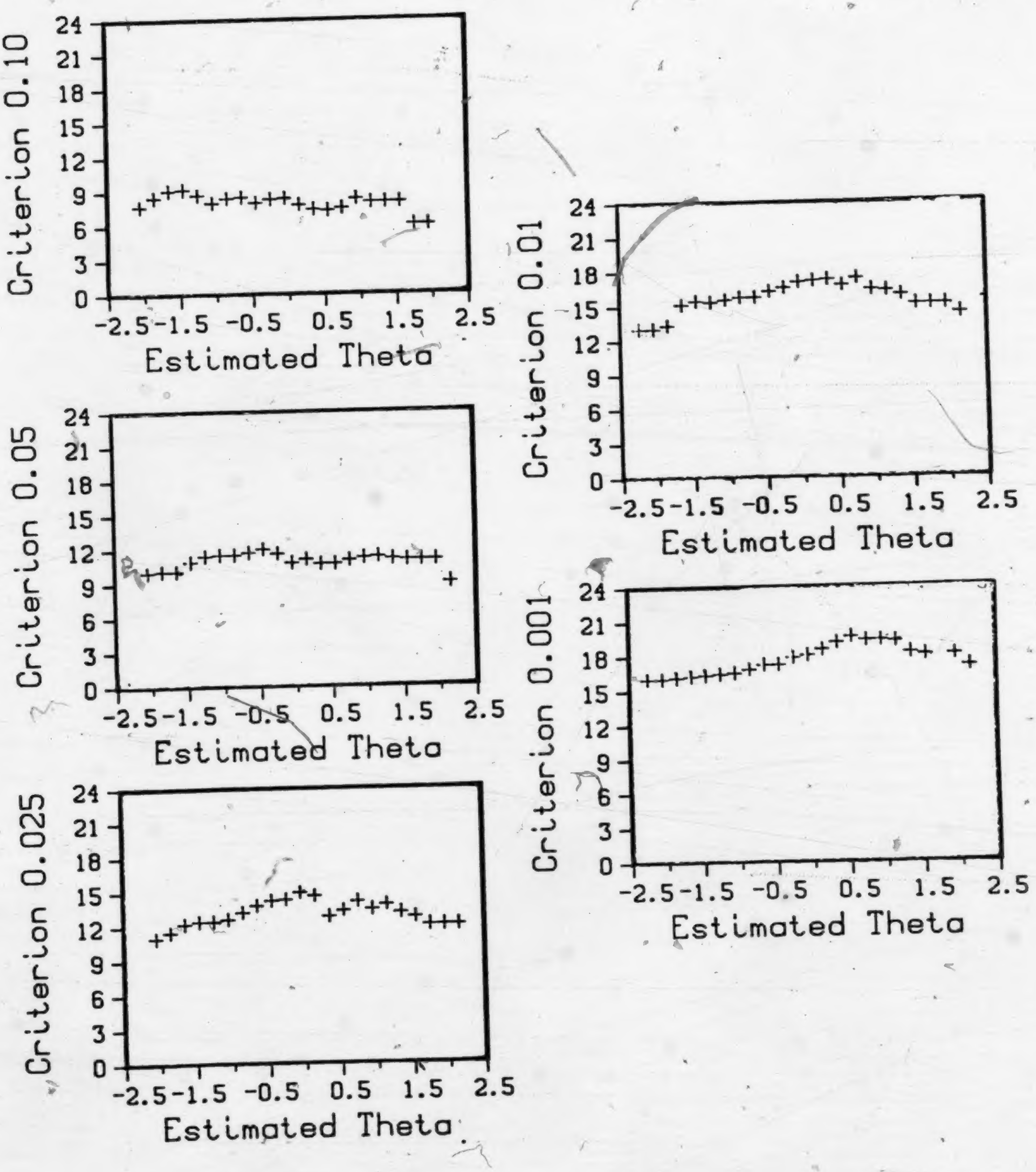


Figure 4.12
Mean Number of Items
vs Estimated Theta - Subtest 2

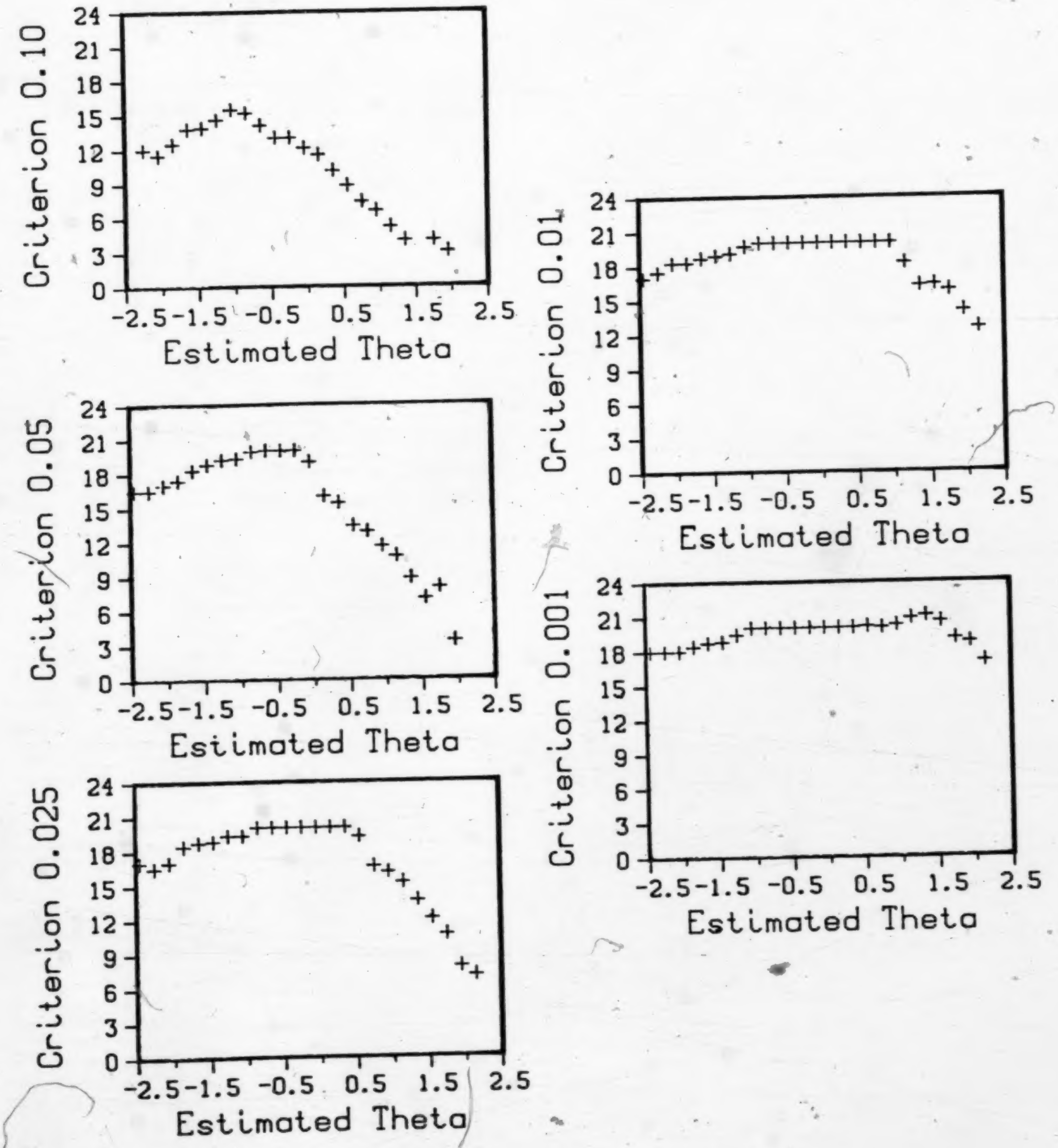


Figure 4.13
Mean Number of Items
vs Estimated Theta - Subtest 3

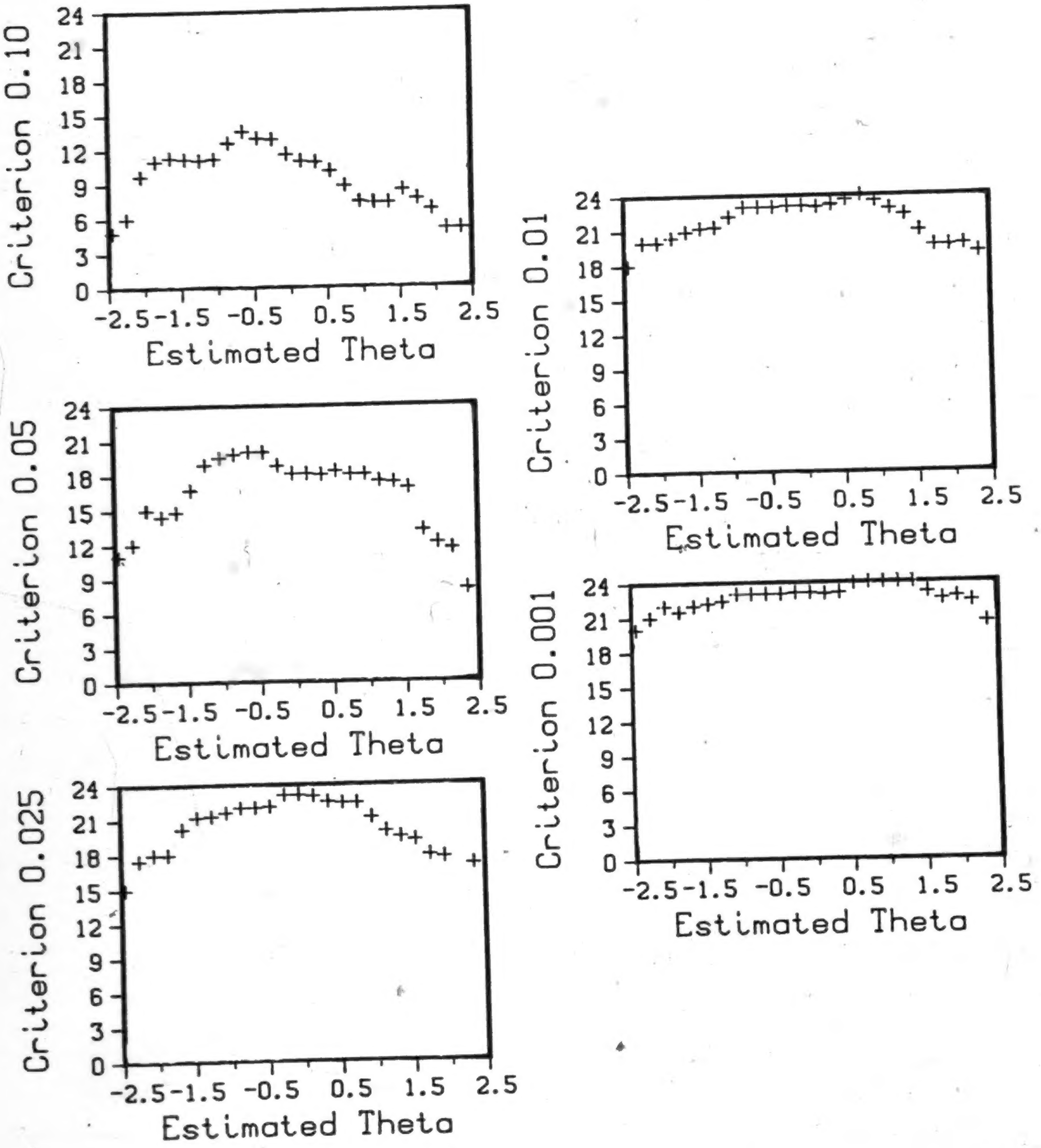
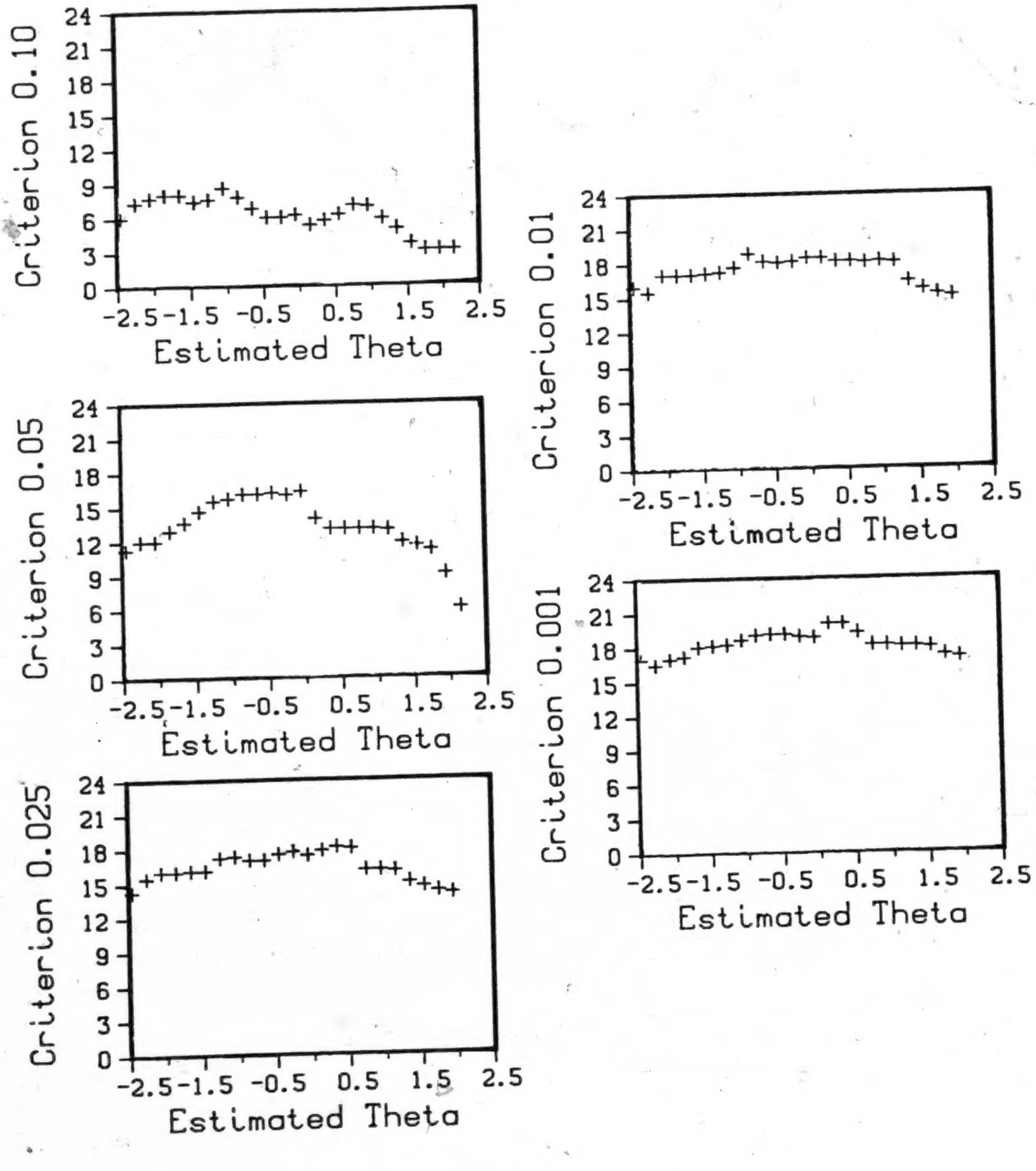


Figure 4.14
Mean Number of Items
vs Estimated Theta - Subtest 4



Consider the simulated tailored testing when criterion C_{00} was used. As noted earlier, this criterion resulted in all items being administered to all examinees. This simulation differs from the conventional administration in that the order in which the items were administered was different. Since all items were administered using criterion C_{00}, the amount of information obtained from this simulation is theoretically greater, given Θ, than the amount obtained from any of the other simulations (*i.e.* those for the other criterion levels).

In order to examine the degree of information loss attributable to item reduction through tailored testing, the average test information curve corresponding to criterion level C_{00} was compared to the average test information curves corresponding to the remaining criterion levels. The comparison was made for each of the four subtests by taking the ratio of the averaged test information value of one curve, say A, at a given value of Θ and dividing it by the averaged test information value of another curve, say B, at the corresponding value of Θ. This ratio was plotted as a function of Θ.

These curves are presented in Figures 4.15 through 4.18. In all cases the ratios were taken between the averaged test information curves under criteria C_{10}, C_{05}, C_{025}, C_{01}, and C_{001} (numerators) and the averaged test information curve corresponding to the simulation using C_{00} (denominator) as the termination criterion. Due to the

Figure 4.15
Observed Efficiencies
For Subtest 1

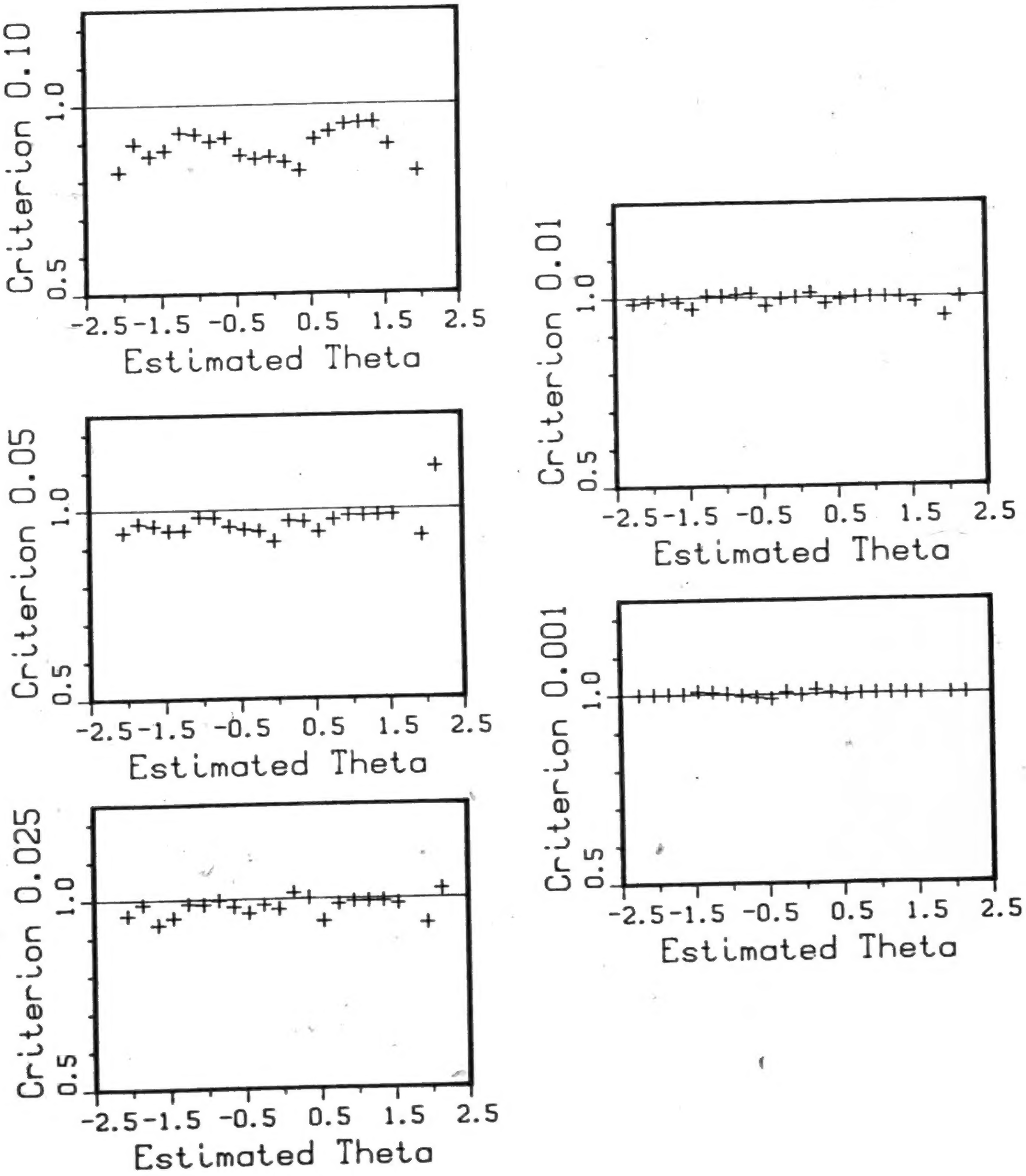


Figure 4.16
Observed Efficiencies
For Subtest 2

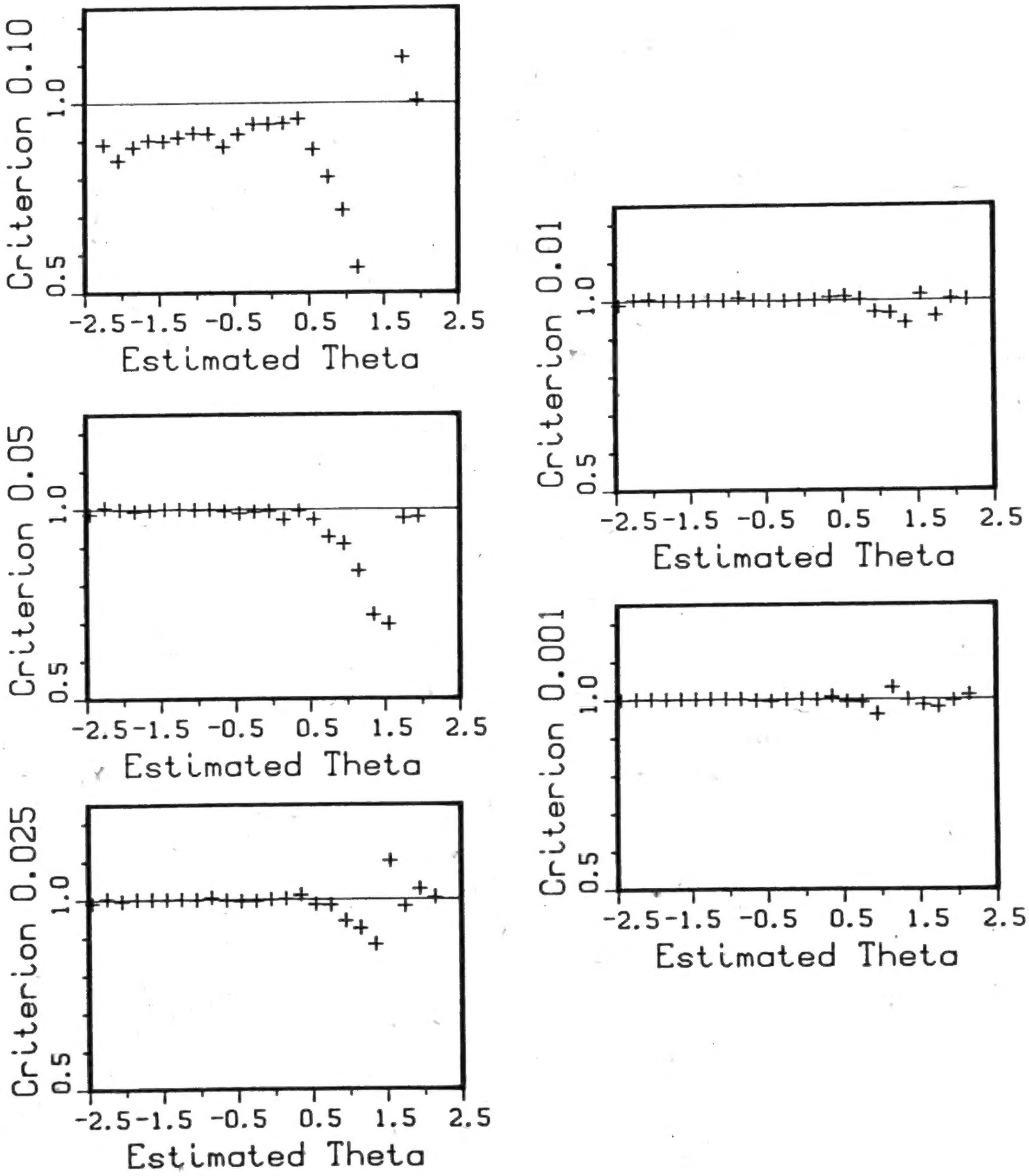


Figure 4.17
Observed Efficiencies
For Subtest 3

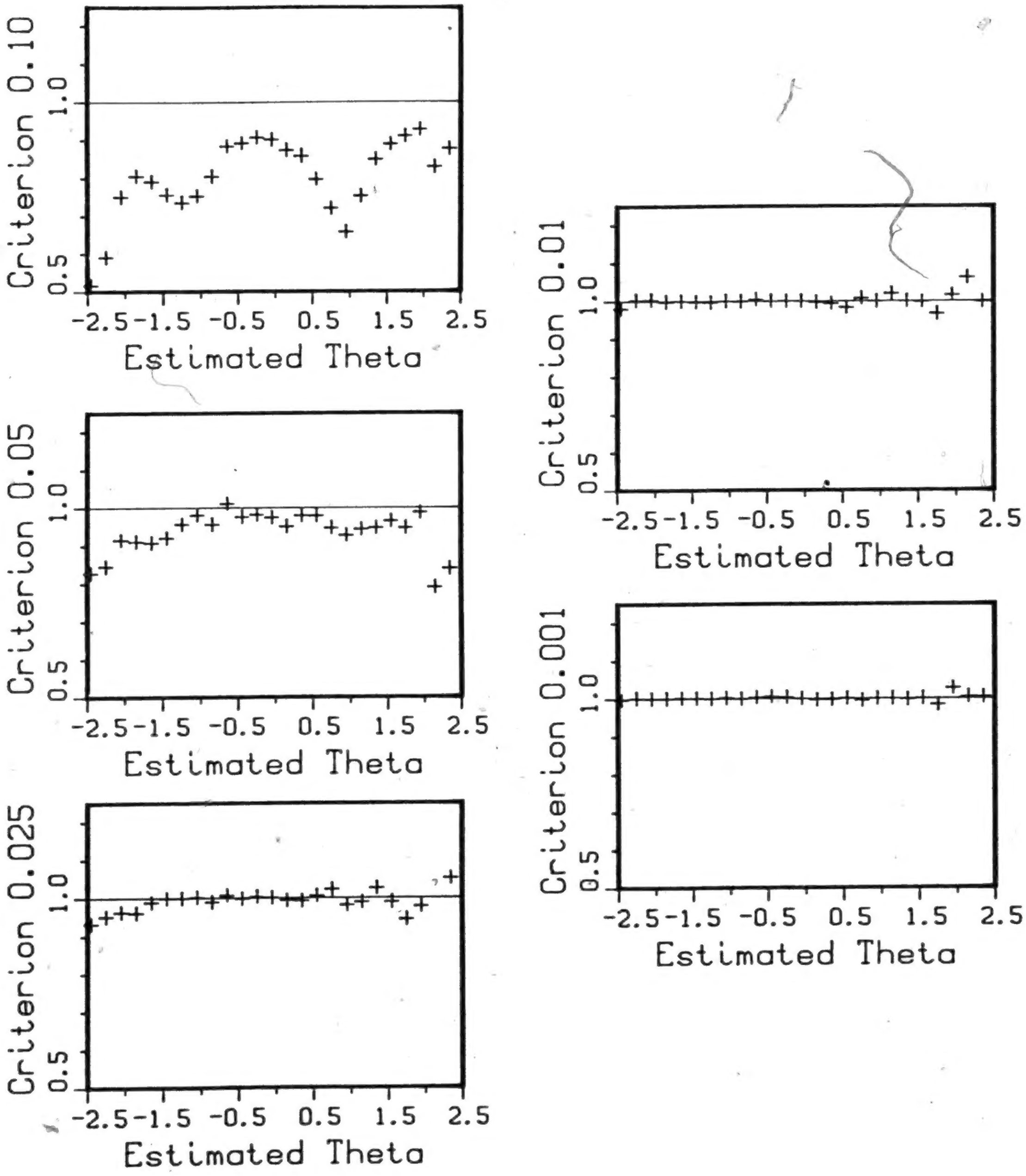
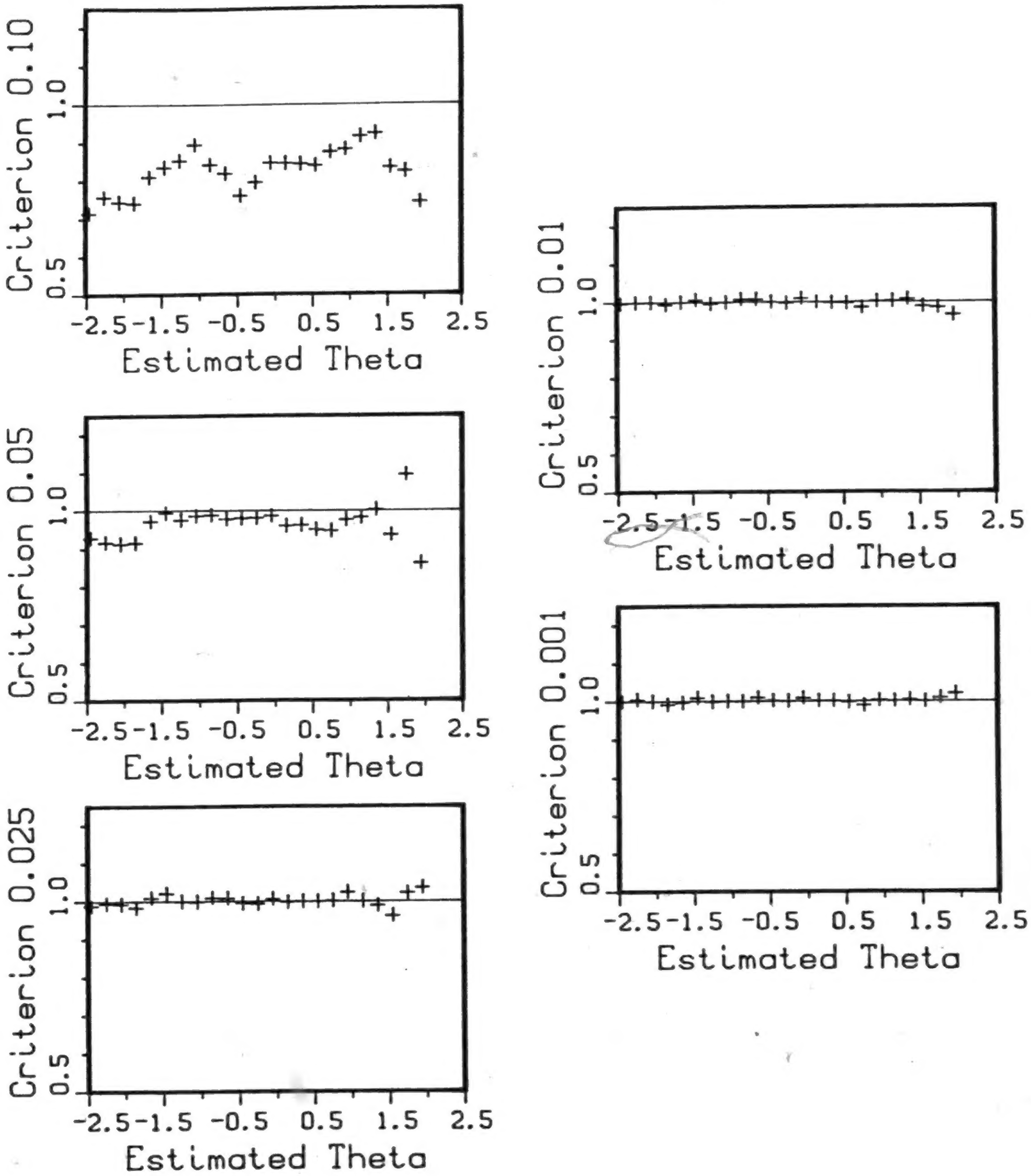


Figure 4.18
Observed Efficiencies
For Subtest 4



discontinuous nature of the observed Θ values, the Θ scale was divided into intervals 0.20 Θ units in width. The averaged test information values found within each interval were averaged to obtain a value for each interval.

At first glance there is an apparent inconsistency between theory and practice. The inconsistency arises because the denominator is theoretically greater than the numerator, yet in some scores the ratios are exceeding 1.00. Closer investigation reveals the cause of the problem. Recall that the ratios are based upon averaged values. These averaged values are themselves based upon average values. The averages are influenced by the variance of the values that are averaged, as well as the number of values on which the average is based. It is therefore considered appropriate to treat any fluctuations of ratios above 1.00 as artifacts of the procedure used and assume that the ratio is equal to 1.00.

It should be mentioned that if the curves to be compared had been actual test information curves (as opposed to averaged test information curves), then Figures 4.15 through 4.18 would represent relative efficiency curves. As they are based upon averaged information values they will be referred to as observed relative efficiency curves.

Interpretation of these graphs is straightforward. In all graphs the observed relative efficiency points are plotted along with a horizontal line at one.[13]

---

[13] This line represents the relative efficiency actually calculated by taking the ratio of the C_{00} averaged test

Observed relative efficiencies less than one (*i.e.* points below the line) indicate the amount of information loss due to the fact that fewer items were administered under a tailoring strategy with a termination criterion different from C_{00}. Values of one indicate equal levels of information from both testing strategies. A visual inspection of these graphs emphasizes the previously discussed result that information increases as more items are administered. With every subtest, as the termination criterion was relaxed, the observed efficiency curve more closely approached the horizontal line representing a relative efficiency of one.

Figures 4.15 to 4.18 show that only minimal amounts of information were lost due to item reduction when the two termination criteria C_{001} and C_{01} were used. Under these criteria, as can be seen from the tables in Appendix C, the greatest loss in averaged relative efficiency was about 6 percent (subtest 1, C_{01}). Excluding this one interval, the averaged relative efficiencies for these two criteria were approximately equal.

From these figures it is clear that the greatest amount of information was lost when the most stringent termination criterion, C_{10}, was used. Losses ranged from 5 to 53 percent with the average being between 15 and 20 percent. This termination criterion made by far the worst showing in this respect. The remaining two criteria, C_{05} and C_{025}, produced reasonable results. For subtests 1, 3, and 4, C_{025} resulted

---

[13](cont'd)information curve to itself.

in minimal information losses, quite comparable to those obtained under C_{01} and C_{001}. For subtest 2, this was also the case except for the interval between $\Theta=1.05$ and $\Theta=1.50$.

Consideration of the results for averaged relative efficiency suggests that the loss of information due to tailored testing was minor except for one termination criterion, C_{10}. In fact, it was evident that the amount of information lost was attributable to the severity of the termination criterion. For all practical purposes, the information curves obtained under tailored testing using the three lowest termination criteria produced information curves comparable to those produced under conventional testing. It seems safe to conclude that the use of a tailoring strategy using reasonable termination criteria did not seriously reduce the information provided by the test.

As the termination criterion relaxes, causing a larger number of items to be administered, the efficiency analysis has shown that the estimate of ability becomes more precise. In terms of the six termination criteria, ability estimates derived under the criterion C_{00} are the most precise.

To investigate the variability of ability estimates across criteria, five difference scores were calculated for each individual by subtracting the ability estimates derived under termination criteria C_{10}, C_{05}, C_{025}, C_{01}, and C_{001} from the estimate derived under C_{00}. Table 4.10 reports the mean and standard deviation for each distribution of difference scores.

Subtest	C_{10}	C_{05}	C_{025}	C_{01}	C_{001}
1	-0.01	0.00	0.10	0.00	0.00
(σ)	(0.21)	(0.14)	(0.10)	(0.72)	(0.32)
Δ	1.11	0.98	0.96	0.96	0.22
2	0.00	0.01	0.00	0.00	0.00
(σ)	(0.19)	(0.11)	(0.09)	(0.80)	(0.03)
Δ	1.34	1.32	1.38	1.41	0.24
3	0.10	0.00	0.00	0.00	0.00
(σ)	(0.22)	(0.11)	(0.06)	(0.05)	(0.01)
Δ	1.63	0.58	0.58	0.50	0.11
4	-0.06	-0.03	-0.02	-0.01	-0.05
(σ)	(0.29)	(0.20)	(0.16)	(0.13)	(0.57)
Δ	1.78	1.60	1.43	1.20	1.84

Table 4.10 - Mean, Standard Deviations, and Maximum Deviation of Tailored Testing Ability Estimates From C_{00}

Two patterns become clear. First, all mean difference scores are close to zero. (In several cases a mean difference score of zero was observed.) Second, aside from the small variances, in all but one case did the variance decrease as the termination was relaxed. Taken together, the mean difference scores and the decreasing variability suggested convergence of the tailored testing ability estimates.

Also reported in Table 4.10 are the maximum difference scores, given in absolute terms. Relatively small values were reported for subtests 1, 2, and 3 but much larger values corresponded to subtest 4. Scatter plots revealed that these values were overestimated as a result of 2 outliers and that with the exclusion of these two points the general pattern of minimal decreasing differences was evident. The outliers were clearly the cause of the relatively large variances of subtest 4 in comparison with subtests 1, 2, and 3. Evidence of outliers was not found in relation to the remaining three subtests.

4.3 Evaluation of Inter-Subtest Branching Strategy

An inter-subtest branching strategy (outlined in section 3.5.3.2) was used in this study to refine the initial ability estimates required for entering subtests 2 through 4. It was expected that if the branching strategy was working properly the number of items administered would be less than if the strategy was not used. Not using the strategy would mean that the initial ability estimate for every examinee would be the same for subtests 2, 3, and 4 as for subtest 1.

The basic strategy used in this study was that proposed by Brown and Weiss (1977) and evaluated favourably by Gialluca and Weiss (1979). As both of these studies involved classroom achievement tests, the present study represented an extension to ability tests.

The effectiveness of the strategy was evaluated by applying the simulated tailored testing procedure to the validation sample data for all subtests, with the initial entry ability set to 0.0 and the initial prior variance set to 1.0 in each case. Recall that these were the same values used on entry to subtest 1 during the simulations reported in the previous sections of this chapter.

The mean number of items administered for each subtest under each criterion without the inter-subtest branching strategy are presented in Table 4.11. The means associated with the simulation that used the inter-subtest branching strategy are found in Table 4.9 (p. 96). Since the strategy was used only to enter subtests 2, 3, and 4, the results for subtest 1 are identical in both tables.

The reduction in the number of items administered under the inter-subtest branching strategy is the difference between the corresponding means in Tables 4.9 and 4.11. For all subtests under all criterion levels a small mean reduction was found. In terms of the total test, the reduction in mean test length ranged between 1.0 percent and 2.7 percent, depending on the termination criterion. This degree of shortening corresponds to a reduction of about one item. These results were somewhat disappointing as Gialluca and Weiss (1979) had found up to a 5 percent reduction attributable to the inter-subtest branching strategy. Still, on the basis of the consistency rather than the magnitude of the reduction, the strategy can be deemed successful. Similar findings were observed when the modes and medians (as opposed to means) were examined.

Subtest	C	C_{10}	C_{05}	C_{025}	C_{01}	C_{001}
1 (σ)	20	7.99 (0.83)	11.13 (0.78)	13.50 (1.10)	16.18 (1.02)	17.82 (1.17)
2 (σ)	21	11.39 (3.30)	16.79 (3.82)	18.71 (2.09)	19.70 (0.68)	19.98 (0.39)
3 (σ)	24	11.44 (1.81)	18.72 (1.44)	22.04 (1.00)	22.91 (0.70)	23.31 (0.47)
4 (σ)	21	6.50 (1.24)	14.96 (1.84)	17.46 (1.09)	18.21 (0.87)	19.18 (0.78)
Total (σ)	86	37.32 (5.57)	61.60 (6.00)	71.72 (3.58)	77.02 (1.89)	80.31 (1.66)

Table 4.11 - Mean and Standard Deviations of the Number of Items Administered Without an Inter-Subtest Branching Strategy

4.4 Increased Precision

Tailored testing is claimed to increase the precision of tests by increasing the discrimination among examinees. If this claim is valid, the range of ability estimates should be greater under tailored testing than under conventional testing. Table 4.12 presents the minimum, maximum, and ranges of estimated abilities obtained under the Bayesian rescoring of the conventional subtests (C), and those obtained through the various tailored testing simulations. The results for subtests 2,3, and 4 generally substantiated the claim that tailored testing would increase the range of ability estimates. The range of estimates for subtest 1, however, were somewhat narrower under tailored than conventional testing.

In the majority of cases, as the termination criterion was relaxed, the range increased. It seemed that the more items administered the wider the range of ability estimates. There did, however, seem to be a point of diminishing returns. The biggest increases in range occured across all subtests when the termination criterion was reduced from 0.10 to 0.05. The next largest increase was found when the termination criterion was reduced from 0.05 to 0.025. As the termination criterion was further reduced, range increases were noted but they were certainly not of a magnitude similar to those already mentioned.

Examining the maximum values (*i.e.* the largest ability estimates) in Table 4.12, it was noted that 23 of the 24

Subtest	C	C_{10}	C_{05}	C_{025}	C_{01}	C_{001}	C_{00}
1 max.	2.01	1.96	2.06	2.11	2.14	2.15	2.15
min.	-2.82	-2.09	-2.09	-2.13	-2.17	-2.16	-2.16
range	4.83	4.05	4.15	4.24	4.31	4.31	4.31
2 max.	1.59	1.97	2.01	2.10	2.15	2.17	2.17
min.	-2.95	-2.73	-2.95	-2.96	-2.98	-2.99	-3.00
range	4.54	4.70	4.96	5.06	5.13	5.16	5.17
3 max.	2.09	2.28	2.34	2.38	2.41	2.41	2.41
min.	-3.02	-2.53	-2.99	-3.16	-3.20	-3.23	-3.22
range	5.11	4.81	5.33	5.54	5.61	5.64	5.63
4 max.	1.83	2.09	2.22	1.99	2.00	2.02	2.02
min.	-3.00	-2.87	-2.91	-3.04	-3.00	-3.02	-3.01
range	4.83	4.96	5.13	5.03	5.00	5.04	5.03

Table 4.12 - Maximum, Minimum, and Range
Values of Ability Estimates

maximums corresponding to the tailored simulations were greater than their corresponding conventional maximums. Examining the minimum values revealed that only 13 out of the 24 tailored minimum values were less than their conventional counterparts. This suggested that the gains in discriminability were at the upper extreme of the ability scale, not at the lower. As the raw score distribution was negatively skewed (*i.e.* skewed to the left) this result seemed reasonable.

To illustrate the gain in discrimination through tailored testing that was realized for individual examinees, four subjects were chosen, two from the upper end of the ability continuum and two from the lower end. Their raw scores together with their ability estimates from tailored testing are presented in Tables 4.13 and 4.14.

Consider first the data for the subjects (A and B) located at the lower extreme of Θ (Table 4.13). In terms of raw scores the subjects are equal. Rescoring the conventional test (*C*), however, results in a 0.07 difference between their ability estimates. As the tailored testing termination criterion approached zero, the difference between the estimates increased until a final difference of 0.15 resulted when all items were administered. The final difference was more than twice that for the rescored conventional test (*i.e. C*).

Similar results are evident for the examinees at the upper end of the ability continuum (Table 4.14). The

S	R	C	C_{10}	C_{05}	C_{025}	C_{01}	C_{001}	C_{00}
A	7	-2.21	-2.19	-2.29	-2.30	-2.32	-2.34	-2.34
B	7	-2.14	-2.08	-2.19	-2.18	-2.18	-2.19	-2.19
Diff	0	0.07	0.11	0.11	0.12	0.14	0.15	0.15

Table 4.13 - Differences Between Ability Estimates of Two Subjects at the Lower End of the Ability Scale

subjects were of equal ability as judged by their raw scores, but as the tailored tests' termination criterion tended toward zero, the spread between their ability estimates increased. Note that in this case the rank order of the subjects interchanged from tailored testing simulations with a stringent termination criterion to tailored testing simulations with a relaxed criterion.

S	R	C	C_{10}	C_{05}	C_{025}	C_{01}	C_{001}	C_{00}
C	20	1.26	1.20	1.11	1.14	1.14	1.45	1.45
D	20	1.11	1.22	1.16	1.19	1.20	1.21	1.21
Diff	0	0.15	0.02	0.05	0.05	0.06	0.24	0.24

Table 4.14 - Differences Between Ability Estimates of Two Subjects at the Upper End of the Ability Scale

4.5 Conclusions

The previous comparison of adaptive and conventional methods of testing revealed exceedingly strong correlations between corresponding subtest scores. At the same time, significant reductions were achieved in the number of items that would be administered. This reduction was shown to have a minimal negative effect on the precision of measurement.

5. Discussion

5.1 Summary

This study was motivated by a statement made by Vern Urry(1977) in one of the earlier articles that introduced the concept of tailored testing to the non-specialist. He claimed that "if tailored testing is to have immediate application, it must use existing test items." (p. 184) From this came the idea to build a series of item pools using an existing test and to investigate the behaviour of a tailored test based on these pools. This study was different from other work in the area in that the item pools used were not specifically designed for tailored testing. The hope was that applications of tailored testing would be stimulated by demonstrating that specially designed item pools are not necessary.

The problem was to determine whether or not the application of a tailoring strategy to an existing, intact test would yield results comparable to those obtained through conventional testing methods. A real-data simulation was carried out using item response data from the verbal battery of the Canadian Cognitive Abilities Test, level F. Raw item response data for this test were rescored under a Bayesian scoring algorithm and were then rescored again under a simulated tailored testing procedure that also involved the Bayesian scoring algorithm. The two testing procedures were compared in terms of correlations between ability estimates, test length, and information I(Θ). These comparisons were made for each of six different tailored testings, which differed in the termination criterion that was applied.

It was found that for each subtest, the correlations between the ability estimates from the conventional testing procedure and the ability estimates from the simulated tailored testing procedure were very high, regardless of the termination criterion. The tailored testing strategy was found to require fewer items than the conventional procedure, although some precision of measurement was lost. The size of the reduction in number of items and extent of the loss in precision of measurement was found to be dependent on the stringency of the criterion used to terminate the tailored testing. These findings were in agreement with those of other studies reported in the literature.

On the basis of this study it was concluded that imposing a tailored testing strategy upon the verbal battery of the Canadian Cognitive Abilities Test, level F had very small negative effects on the resulting ability estimates.

5.2 Discussion and Implications

The results of this study generally parallel those obtained in earlier work (Brown & Weiss, 1977; Gialluca & Weiss, 1979), which used achievement tests. Perhaps the most striking result of the present study was that tailored testing could reduce test length by as much as 55%, hence reducing the time spent in testing. This reduction is larger than those reported elsewhere in the literature, but the termination criterion used in this case was far more stringent than those used in previous studies. Moreover, the most stringent criterion resulted in a much lower level of information than for the conventional test (with all items administered). Reductions in length under the less restrictive termination criteria were more in line with the reductions reported in previous research. The actual amount by which test length was reduced was found to be dependent on the termination criterion.

It was important in this study to show the potential of tailored testing for reducing the time spent in test taking. The economics of education are no different from those of commerce in that time means money. Testing and evaluation is on the rise and more educational time is being spent on testing.

As education attempts to keep step with current technological advances, increasing numbers of computers are finding their way into the classroom. With the development of some software, tailored testing could be implemented on these computers, thus, perhaps, reducing the cost of testing by reducing the time spent on testing.

If tailored testing is to be used in the classroom, it must be adapted to the micro-computer environment. There are three reasons why micro-computers are preferred to mini or mainframe computers. The first reason is that micro-computers are presently being used in the schools; bringing in other kinds of computer hardware would be an expensive duplication. Second, the cost of mini-computers (and especially mainframes) with terminals is prohibitive. And third, a number of users (say 25) would drastically reduce the speed of a mini-computer to the point that any savings in testing time due to teacher testing would be lost through computer processing delays.

The basic characteristics of a micro-computer that could be used for tailored testing are these: 128K capacity; 80 column screen width; dual floppy disk drives; and a computer language such as FORTRAN (*i.e.* the FORTRAN '77 package). At least two popular brands of micros can satisfy these requirements, the IBM PC and the Apple IIe micro.

At least two scenarios are possible when considering hardware configurations practical for tailored testing in a school environment. The simplest is a collection of a number of micro-computers. To administer a test two diskettes would be required. One diskette would contain the data bases and the software necessary to administer and score the test. The second diskette would record response information, ability estimates, and any other information obtained through testing. Depending upon the testing situation the second diskette could record information from a group of examinees or be assigned to a single examinee and become a record of his test results.

A major problem with this kind of set-up would be the length of time required to test a large number of examinees. This would be purely a function of the number of micro-computers available for testing.

If many micro-computers are available a second scenario is more practical than the one just discussed. In this situation the micros would be linked together to form a local area network and would be controlled by a file server device. This device would down load the necessary data bases and necessary software and might also act as a storage/output device for the results of testing. This approach would be more expensive, due to the additional hardware requirements, but it could prove to be more useful, both for instruction as well as for testing.

To take tailored testing into the classroom will involve initial cost (eg. hardware), but as computers become more commonplace in the classroom computerized testing may prove to be a worthwhile application of computers in education. Bejar, Weiss, and Gialluca (1977) make an interesting comment in this regard.

> In order to exploit the advantages of adaptive testing ... it will be necessary to build a closer psychometric interface between instruction and testing. Reduction in testing time ... is meaningless if the [sole] result is ... early dismissal from examinations. Rather what is needed is to link adaptive testing with an adaptive

> instructional context, so that reductions in testing time can be used in increased instructional activity. (p. 26)

Even though this remark was made in the context of achievement testing it is relevant to other forms of testing, including the testing of mental ability. Regardless of type and purpose, testing reduces the number of hours available for instruction. If the number of tests administered cannot be reduced, then methods that will improve the efficiency of testing should be seriously considered for application wherever possible.

A result obtained by Gialluca and Brown (1979) was also obtained in this study. The item parameters, subtest intercorrelations, and regression equations (*i.e.* the inter-subtest branching strategy) determined from data from the calibration sample worked successfully in the simulation using the validation sample. This illustrates IRT's ability to make measurements across groups of people without distorting the characteristics of the items (*i.e.* person free measurement).

Tailored testing is claimed to improve measurement by administering a minimum number of items to each examinee while maintaining a relatively high level of information or, alternatively, a relatively low standard error. Tailored testing is also claimed to increase the discrimination among examinees, especially at the extremes of the ability levels. Each of these claims were supported by the results of this study, however, degree of support for these claims was a function of the criterion level used to terminate the testing procedure.

Consider the claim of item reduction with minimal information loss. It was apparent from Table 4.9 that as the criterion became more stringent, the number of items was reduced. However, Figures 4.15 to 4.18 indicated that as the criterion became more stringent, precision was lost. Clearly, there exists a trade-off between item reduction and information loss. For tailored testing to yield maximum benefits, the termination criterion must allow for significant item reduction while minimizing information loss.

In this study two criterion levels, C_{025} and C_{01}, appeared to achieve this aforementioned objective. Table 4.9 showed that reductions of up to 32% in the number of items could be realized for these levels of the termination criterion while at the same time Figures 4.15 through 4.18 showed that the corresponding loss of information was minimal. More stringent criteria produced greater reductions in the number of items, but also produced greater losses of information.

The second advantage claimed for tailored testing is increased discrimination among examinees. It was demonstrated in section 4.4 that discrimination varies with the stringency of the termination criterion. In most cases the range of ability estimates increased as the criterion became more stringent, but a point of diminishing returns was reached, and the fact that it had been reached was much more obvious than for item reduction and information loss. Examining Table 4.13 and keeping in mind the discussion in the previous paragraph, we see that criterion level C_{025} provided the best demonstration of the advantages of tailored testing. Termination criterion C_{01} also produced acceptable results, but with this criterion more items were administered than for criterion C_{025} with only a marginal improvement in information.

Generalization of the results beyond the verbal battery of the CCAT would be hazardous due to the particular characteristics of the items. Still, the results are sufficiently encouraging that other ability tests or batteries of ability tests should be examined for results comparable to those obtained here.

5.3 Directions For Future Study

Research in several areas is necessary to develop and refine tailored testing. Perhaps the most important area is the development and refining of tailored testing strategies. The development of more efficient intra-subtest branching strategies is necessary if item pools are to be used to best advantage, whether they are designed especially for tailored testing or, as in this study, formed from existing conventional tests. It is likely that there does not exist an item selection strategy that is best for all item pools.

Research in this area should be directed toward defining salient characteristics of item pools, and employing different item selection strategies to then see if the strategy of choice varies with pool characteristics.

Work in this area should not be restricted to intra-subtest branching, but should also be directed at inter-subtest branching. As tailored testing is applied to multi-dimensional tests, inter-subtest strategies must be devised that will take advantage of the intercorrelations of the traits. Since it is rare to find two ability traits that are not correlated to some extent, the objective should be to use the information obtained in measuring one trait to refine the measurement of the other. This was attempted in the present study through the use of regression equations and differential entry points to subtests after the first, but as Gialluca and Brown (1979) point out, such a procedure is vulnerable to errors of measurement in the independent variables.

Related to the problem of selecting an intra-subtest branching strategy is the problem of determining an appropriate termination criterion for the tailored testing strategy used in the subtest. This study showed that the choice of termination criterion could dramatically influence level of information. This result was due to the particular characteristics of the item pools that were used, but it does underline the need for care in selecting criteria.

For the time being, research in the aforementioned areas should be conducted using simulated testing procedures. Two features of simulated tailored testing studies make this approach attractive. First, real data from real examinees is used, not manufactured data. In comparison with Monte Carlo studies, this feature gives simulated studies the advantage of credibility. Second, large data sets are readily available from existing large scale testing programs. Large sample sizes are needed to ensure accurate IRT item parameter estimates and to provide a large validation sub-sample.

Research has begun (Samejima, 1979; Wood, 1983) on polychotomous response models in tailored testing. These models attempt to improve measurement by considering the information in an incorrect choice. Proponents of these models feel that useful information is contained in each distractor. Even though work in this area is still in its infancy, these new models are promising.

As tailored testing moves into the realm of educational testing, the basic IRT assumption of unidimensionality becomes untenable. Studies should be done to determine exactly how and in which way violations of this assumption effect ability estimation. Further work is needed on how to use multidimensional item pools to best advantage.

Tailored testing lends itself quite readily to the area of computer assisted instruction (CAI). Since instructional material is presented via computer terminals, the hardware is already available; thus one of the major drawbacks of tailored testing is not a concern. Some CAI installations presently use conventional testing procedures, and a move toward tailored testing would require software to control the branching and scoring strategies. As tailored testing is developed, it will most likely become part of the CAI environment.

5.4 A Final Word

The results of this study are encouraging. The application of tailored testing to already existing item pools was successfully demonstrated. It does remain to be seen whether similar results can be obtained for other intact tests. There is need for studies similar to this, but for other tests. Should they be as successful as this study, practioners and educators may be encouraged to replace conventional testing with tailored testing procedures.

In some areas, practioners have already recognized the potential of tailored testing. Applications are being made at the University of Minnesota and the U.S. Civil Service Commission. That interest in tailored testing is mounting is evident from the large number of research contracts being funded through the United States military. In Vancouver, Canada, pencil and paper tests are no longer used to assess applicants for a driver's license. Rather, through Telidon (TV information stystem), candidates are given a tailored test that assesses their knowledge of licensing regulations, rules of the road, driving skills, practices and attitudes, vehicle condition, and general background knowledge. Advocates of tailored testing find their dreams fulfilled by such diverse and successful applications as these.

Item response theory has come a long way since its introduction by Lord over three decades ago. At that time, practical applications of IRT were almost unknown. But as computer technology has evolved, the feasibility of tailored testing has also grown. On the basis of the attention it is receiving in the literature, it seems safe to say that the age of tailored testing is near at hand. With careful nurturing, tailored testing will soon revolutionize modern testing.

Bibliography

Barton, M. A., Lord, F. M. *An upper asymptote for the three-parameter logistic item-response model*. Princeton: Educational Testing Service, July 1981.

Bejar, I. I., & Weiss, D. J. *A construct validation of adaptive achievement testing* (Research Report 78-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods, November 1978.

Bejar, I. I., Weiss, D. J. *Computer program for scoring test data with item characteristic curve models* (Research Report 79-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods, February 1979.

Bejar, I. I., Weiss, D. J., & Gialluca, K. A. *An information comparison of conventional and adaptive tests in the measurement of classroom achievement* (Research Report 77-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods, October 1977.

Bejar, I. I., Weiss, D. J., & Kingsbury, G. G. *Calibration of an item pool for the adaptive measurement of achievement* (Research Report 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods, September 1977.

Binet, A. *The development of intelligence in children*. Baltimore: Williams & Wilkins, 1916.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores* (Part 5). Reading, Mass: Addison-Wesley, 1968.

Brown, J. M., & Weiss, D. J. *An adaptive testing strategy for achievement test batteries* (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods, October 1977.

Divgi, D. R. Does the Rasch model really work? Not if you look closely. *Paper presented at the annual meeting of the National Council on Measurement in Education*, Los Angeles, 1981.

Ebel, R. L. *Essentials of educational measurement*. Englewood Cliffs, New Jersey: Prentice-Hall, 1972.

Gialluca, K. A., & Weiss, D. J. *Efficiency of an adaptive inter-subtest branching strategy in the measurement of classroom achievement* (Research Report 79-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods, November 1979.

Green, B. F. Comments on tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance*. New York: Harper & Row, 1970.

Green, M. S., & Divgi, D. R. The invariance of parameter estimates in three latent models. *Paper presented at the annual meeting of the National Council on Measurement in Education*, Los Angeles, 1981.

Hambleton, R. K. Latent trait models and their applications. In R. Traub (Ed.), *Methodological developments: New directions for testing and measurement* (No. 4). San Francisco: Jossey-Bass, 1979.

Hambleton, R. K. *Latent ability scales: interpretations and uses*. A report prepared for the U. S. Air Force Human Resources Laboratory, Washington, 1980.

Hambleton, R. K., Cook, L. L. Latent trait models and their use in the analysis of educationaal test data. *Journal of Educational Measurement*, 1977, *14*(2), 75-96.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D., & Gifford, J. A. Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Research*, 1978, *48*(4), 467-510.

Harley, D. D. *The Henmon-Nelson: Computerized*. Unpublished Master of Arts thesis, University of British Columbia, 1979.

Hattie, J. *An empirical study of various indices for determining unidimensionality*. 1983 - Manuscript submitted for publication. (a)

Hattie, J. *A review of methods for assessing unidimensionality of tests and items*. 1983 - Manuscript submitted for publication. (b)

Hedl, J., O'Neil, H. F, & Hansen, D. N. *Computer based intelligence testing*. Florida State University, 1971. (ERIC Document No. Ed0605981)

Jensema, C. T. Bayesian tailored testing and the influence of item bank characteristics. *Applied Psychological Measurement*, 1977, *1*, 111-120.

Jensema, C. T. The validity of Bayesian tailored testing. *Educational and Psychological Measurement* , 1974, *34*, 757-766.

Kerlinger, F. N. *Foundations of behavioral research* (2nd Ed.). New York: Holt, Rinehart and Winston, 1973.

Kreitzberg, C. B., Stocking, M. L., & Swanson, L. Computerized adaptive testing: Principals and directions. *Computers and Education*, 1978, *2* (4), 319-329.

Koch, W. R., & Reckase, M. D. *Problems in application of latent trait models to tailored testing* (Research Report 79-1). Columbia: University of Missouri, Tailored Testing Research Laboratory, Educational Psychology Department, September 1979.

Kuder, G. F. , & Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, *2*(3), 151-160.

Lord, F. M. A theory of test scores. *Psychometric Monographs #7*, 1952.

Lord, F. M. An application of confidence intervals and maximum liklihood to the estimation of an examinee's ability. Psychometrika, 1953, 18, 57-75. (a)

Lord, F. M. The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 1953, 13, 517-548. (b)

Lord, F. M., & Novick, M. Statistical theories of mental test scores. Reading, Mass: Addison-Wesley, 1968.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper & Row, 1970.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39(2), 247-264.

Lord, F. M. A broad-range tailored test of verbal ability (Research Bulletin RB-75-5). Princeton, N. J.: Educational Testing Service, February 1975.

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Erlbaum, 1980.

McKinley, R. L., & Reckase, M. D. A successful application of latent trait theory to tailored achievement testing (Research Report 80-1). Columbia: University of Missouri, Tailored Testing Research Laboratory, Educational Psychology Department, February 1980.

McKinley, R. L., & Reckase, M. D. A comparison of a Bayesian and a maximum likelihood tailored testing procedure (Research Report 81-2). Columbia: University of Missouri, Tailored Testing Research Laboratory, Educational Psychology Department, June 1981.

Miyazaki. I. China's examination hell (C Schirokauer, trans.). New York: Weatherhill, 1976. (Originally published, 1963.)

Nelson, L. R. LERTAP. Colorado: University of Colorado, Laboratory of Educational Research, 1974.

Novick, M. R. Bayesian methods in psychological testing (Research Report RB-69-31). Princeton, N. J.: Educational Testing Service, April 1969.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Samejima, F. A. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1(2), 233-247.

Samejima, F. A. A new family of models for the multiple choice item (Research Report 79-4). Knoxville: University of Tennessee, Department of Psychology, December 1979.

Samejima, F. A. Is Bayesian estimation proper for estimating the individual's ability? (Research Report 80-3). Knoxville, Tennessee: University of Tennessee, Department of Psychology, July 1980.

Sax, G. Principles of educational measurement and evaluation. Belmont, Calif: Wadsworth, 1974.

Spearman, C. The theory of two factors. Psychological Review, 1914, 21, 101-115.

Stevens, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), Handbook of experimental psychology . New York: John Wiley, 1951.

Swaminathan, H. Parameter estimation in item response theory models. In R. Hambleton (Ed.), *Applications of item response theory*. Vancouver: Educational Research Institute of British Columbia, 1983.

Sympson, J. B. Estimation of latent trait status in adaptive testing procedures. In D. J. Weiss (Ed.), *Applications of computerized adaptive Testing* (Research Report 77-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods, March 1977.

Thompson, E. G., & Weiss, D. J.*Criterion-related validity of adaptive testing strategies*. (Research Report 80-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods, June 1980.

Thorndike, E. L. *Mental and social measurements* (2nd Ed.). New York: Teachers College, Columbia University, 1919.

Thorndike, R. L., & Hagen, E. *Canadian cognitive abilities test*. Toronto: Thomas Nelson, 1974.

Traub, R., & Wolfe, R. Latent trait theories and the assessment of educational achievement. In D. Berliner (Ed.) *Review of research in education* (Vol. 9). American Educational Research Association, 1981.

Urry, V. W. A Monte Carlo investigation of logistic mental test models (Doctoral dissertation, Purdue University, 1971). *Dissertation Abstracts International*, 1971, *31*, 6319B. (Abstract)

Urry, V. W. Tailored testing: a successful application of latent trait theory. *Journal of Educational Measurement*, 1977, *14*, 181-196.

Warm, T. *A primer of item response theory* (940269). Oklahoma City: U. S. Coast Guard Institute, December 1978. (NTIS No. AD-A063072)

Weiss, D. J., & Betz, N. E. Ability measurement: Conventional or adaptive? (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods, November 1973. (NITS No. AD 757788)

Weiss, D. J., & Davidson, M. L. Review of test theory and methods (Research Report 81-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods, January 1981.

Wood, G. T. Computer-aided item development: An application of Samejima's new family of models. In R. Hambleton (Ed.), Applications of item response theory. Vancouver: Educational Research Institute of British Columbia, 1983.

Wood, R. L., & Lord, F. M. A user's guide to LOGIST. Research Memorandum 76-4. Princeton, NJ: Educational Testing Service, 1976.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, NJ: Educational Testing Service, 1976.

Wright, R. D., & Stone, M. H. Best test design. Chicago: MESA, 1979.

6. Appendix A

The three item response parameters for each item used in the study are listed by subtest in tables 6.1 through 6.4. Along with these parameters, three classical test theory item statistics are presented: difficulty index, point-biserial correlation coefficient, and biserial correlation coefficient. The difficulty index is simply the proportion of examinees who gave a correct response to the item. The two correlation coefficients reflect the direction and magnitude of the relationship between the binary (0,1) item response and the total test score. These correlation coefficients provide a measure of item discrimination.

Subtest A - Vocabulary						
Item	*a*	*b*	*c*	Diff[1]	P-Bis[2]	Bis[3]
1	*0.598*	*-4.425*	*0.170*	*98.1*	*0.16*	*0.32*
2	0.664	-3.283	0.170	95.8	0.25	0.48
3	0.511	-1.098	0.170	74.2	0.38	0.51
4	*0.270*	*-5.531*	*0.170*	*93.2*	*0.18*	*0.31*
5	0.123	-0.738	0.170	61.6	0.24	0.31
6	*0.131*	*-9.005*	*0.170*	*89.9*	*0.15*	*0.25*
7	0.502	-2.930	0.170	91.2	0.28	0.46
8	*0.936*	*-3.775*	*0.170*	*98.7*	*0.20*	*0.34*
9	1.038	0.054	0.170	57.2	0.48	0.60
10	0.843	-2.127	0.170	91.4	0.34	0.57
11	*0.173*	*-6.904*	*0.170*	*89.9*	*0.17*	*0.28*
12	2.587	-1.092	0.170	85.2	0.43	0.65
13	0.752	-0.826	0.170	73.9	0.41	0.56
14	0.228	-0.197	0.170	59.8	0.30	0.38
15	0.259	2.621	0.170	37.7	0.26	0.33
16	0.987	-2.572	0.170	95.4	0.31	0.58
17	0.287	-2.528	0.170	80.0	0.27	0.39
18	0.883	-0.615	0.170	71.4	0.44	0.58
19	2.357	0.403	0.225	51.4	0.46	0.58
20	2.451	1.406	0.103	19.9	0.34	0.49
21	0.428	1.900	0.170	35.8	0.30	0.38
22	0.710	0.419	0.170	50.6	0.43	0.54
23	0.445	-0.493	0.170	65.1	0.38	0.50
24	2.717	0.783	0.140	35.0	0.48	0.61
25	2.442	1.639	0.200	26.1	0.26	0.35
μ	1.061	-0.464	0.169	62.9	0.35	0.49
σ	0.896	1.670	0.022	23.6	0.08	0.10

[1] Diff - proportion of examinees who responded correctly

[2] P-Bis - point biserial correlation between the binary (0,1) item response and the total test score.

[3] Bis - biserial correlation between the binary (0,1) item response and the total test score.

Note: Items shown in *italics* were removed from the study. The mean (μ) and standard deviation (σ) were calculated using only those items used in the simulation.

Table 6.1 Extended Item Parameters for Subtest A

Subtest B - Sentence Completion						
Item	*a*	*b*	*c*	Diff[1]	P-Bis[2]	Bis[3]
1	*0.176*	*-13.746*	*0.130*	*98.6*	*0.08*	*0.16*
2	*0.476*	*-4.266*	*0.130*	*96.4*	*0.21*	*0.39*
3	0.454	-3.245	0.130	92.0	0.27	0.46
4	0.711	-1.704	0.130	86.0	0.41	0.62
5	0.668	-1.338	0.130	80.5	0.44	0.62
6	0.617	-3.129	0.130	95.0	0.27	0.50
7	*0.763*	*-3.623*	*0.130*	*98.2*	*0.22*	*0.44*
8	*0.840*	*-4.334*	*0.130*	*99.4*	*0.14*	*0.27*
9	0.685	-2.159	0.130	90.1	0.36	0.59
10	0.412	-2.638	0.130	86.6	0.30	0.46
11	0.883	-2.312	0.130	93.9	0.36	0.65
12	0.474	-2.078	0.130	84.2	0.35	0.52
13	0.668	-2.994	0.130	95.2	0.28	0.52
14	0.838	-1.737	0.130	88.3	0.40	0.63
15	1.118	-0.521	0.130	70.5	0.51	0.67
16	1.202	-1.643	0.130	90.7	0.43	0.71
17	0.775	-1.514	0.130	84.7	0.44	0.65
18	0.778	-1.094	0.130	78.7	0.45	0.63
19	0.630	-0.293	0.130	62.9	0.45	0.58
20	0.767	0.295	0.130	51.4	0.47	0.59
21	0.437	-3.158	0.130	90.9	0.28	0.46
22	0.403	-1.227	0.130	72.9	0.37	0.49
23	2.461	-0.006	0.150	58.1	0.56	0.70
24	2.565	-0.117	0.100	59.3	0.59	0.75
25	3.583	1.945	0.137	20.8	0.25	0.36
μ	1.006	-1.460	0.130	77.8	0.39	0.58
σ	0.830	1.330	0.008	18.5	0.10	0.10

[1] Diff - proportion of examinees who responded correctly

[2] P-Bis - point biserial correlation between the binary (0,1) item response and the total test score.

[3] Bis - biserial correlation between the binary (0,1) item response and the total test score.

Note: Items shown in *italics* were removed from the study. The mean (μ) and standard deviation (σ) were calculated using only those items used in the simulation.

Table 6.2 Extended Item Parameters for Subtest B

Subtest C - Verbal Classification						
Item	*a*	*b*	*c*	Diff[1]	P-Bis[2]	Bis[3]
1	*0.682*	*-4.152*	*0.185*	*98.6*	*0.19*	*0.38*
2	*0.244*	*-6.618*	*0.185*	*94.7*	*0.17*	*0.31*
3	0.357	-2.180	0.185	81.5	0.30	0.43
4	0.621	-3.024	0.185	94.7	0.27	0.49
5	0.423	0.357	0.185	54.5	0.37	0.47
6	0.665	-2.844	0.185	94.5	0.28	0.51
7	0.456	-2.882	0.185	90.2	0.27	0.44
8	1.136	-2.711	0.185	97.4	0.28	0.55
9	0.443	-1.811	0.185	81.3	0.33	0.47
10	2.841	0.064	0.192	58.7	0.47	0.60
11	1.232	-3.356	0.185	99.1	0.24	0.46
12	*0.406*	*-3.579*	*0.185*	*92.3*	*0.23*	*0.40*
13	2.701	-1.121	0.185	86.8	0.41	0.63
14	0.742	-1.025	0.185	77.7	0.42	0.58
15	0.685	-0.731	0.185	72.5	0.41	0.54
16	0.482	-1.049	0.185	73.8	0.36	0.48
17	0.338	-1.439	0.185	74.2	0.31	0.41
18	0.381	-0.433	0.185	64.4	0.34	0.44
19	2.758	1.122	0.114	25.5	0.39	0.53
20	0.632	1.966	0.170	29.4	0.29	0.38
21	0.633	-1.195	0.185	78.8	0.38	0.53
22	*0.092*	*-4.371*	*0.185*	*72.7*	*0.20*	*0.27*
23	0.394	-0.501	0.185	65.5	0.34	0.44
24	0.463	1.299	0.185	42.1	0.34	0.42
25	2.732	1.055	0.209	34.7	0.37	0.48
μ	1.005	-0.973	0.182	70.3	0.34	0.49
σ	0.901	1.553	0.017	22.4	0.06	0.07

[1] Diff - proportion of examinees who responded correctly

[2] P-Bis - point biserial correlation between the binary (0,1) item response and the total test score.

[3] Bis - biserial correlation between the binary (0,1) item response and the total test score.

Note: Items shown in *italics* were removed from the study. The mean (μ) and standard deviation (σ) were calculated using only those items used in the simulation.

Table 6.3 Extended Item Parameters for Subtest C

Subtest D - Verbal Analogies

Item	*a*	*b*	*c*	Diff[1]	P-Bis[2]	Bis[3]
1	0.328	-1.843	0.145	76.1	0.30	0.41
2	0.915	-1.750	0.145	88.6	0.39	0.62
3	0.458	-2.523	0.145	87.0	0.29	0.45
4	0.746	-0.328	0.145	63.9	0.45	0.57
5	0.505	-0.978	0.145	71.8	0.38	0.50
6	0.547	-1.309	0.145	77.2	0.37	0.51
7	0.623	-2.820	0.145	92.9	0.29	0.50
8	0.417	-0.604	0.145	65.2	0.36	0.46
9	0.479	-1.506	0.145	77.9	0.36	0.49
10	0.419	-0.896	0.145	69.0	0.33	0.44
11	0.639	-3.301	0.145	95.5	0.26	0.47
12	*0.117*	*-7.199*	*0.145*	*83.2*	*0.17*	*0.24*
13	0.477	-1.600	0.145	79.0	0.33	0.47
14	0.411	0.540	0.145	50.1	0.35	0.44
15	0.784	0.538	0.145	46.3	0.43	0.54
16	0.802	-0.563	0.145	69.1	0.44	0.58
17	0.240	-2.174	0.145	74.3	0.25	0.34
18	0.708	0.540	0.145	46.8	0.42	0.53
19	0.322	0.465	0.145	52.3	0.30	0.38
20	2.477	-0.011	0.179	59.7	0.51	0.65
21	0.564	-1.471	0.145	79.5	0.38	0.53
22	2.256	-0.367	0.145	69.1	0.52	0.68
23	0.576	0.574	0.145	47.5	0.35	0.44
24	2.907	1.556	0.101	18.0	0.32	0.47
25	0.502	2.681	0.090	19.8	0.29	0.41
μ	0.796	-0.715	0.142	65.69	0.36	0.50
σ	0.702	1.423	0.016	20.23	0.07	0.08

[1] Diff - proportion of examinees who responded correctly

[2] P-Bis - point biserial correlation between the binary (0,1) item response and the total test score.

[3] Bis - biserial correlation between the binary (0,1) item response and the total test score.

Note: Items shown in *italics* were removed from the study. The mean (μ) and standard deviation (σ) were calculated using only those items used in the simulation.

Table 6.4 Extended Item Parameters for Subtest D

7. Appendix B

The following tables contain the number of observations, the average number of items administered, and the average posterior variance per Θ interval from the simulated testings. This data is graphically displayed in Figures 4.7 through 4.14. The following abreviations are used throughout Tables 7.1 to 7.12.

f	number of observations within interval.
n of Items	mean number of items administered within interval.
Post. Var.	averaged posterior (error) variance within interval.

SUBTEST 1							
		Criterion = 0.10			Criterion = 0.05		
Θ Interval		f	n of Items	Post. Var.	f	n of Items	Post. Var.
-3.55	-3.36	0	0.0	0.0	0	0.0	0.0
-3.35	-3.16	0	0.0	0.0	0	0.0	0.0
-3.15	-2.96	0	0.0	0.0	0	0.0	0.0
-2.95	-2.76	0	0.0	0.0	0	0.0	0.0
-2.75	-2.56	0	0.0	0.0	0	0.0	0.0
-2.55	-2.36	0	0.0	0.0	0	0.0	0.0
-2.35	-2.16	0	0.0	0.0	0	0.0	0.0
-2.15	-1.96	3	7.7	0.18	3	10.0	0.17
-1.95	-1.76	10	8.5	0.18	12	10.1	0.18
-1.75	-1.56	8	9.1	0.19	8	10.1	0.18
-1.55	-1.36	11	9.2	0.19	11	10.9	0.19
-1.35	-1.16	21	8.7	0.19	18	11.4	0.18
-1.15	-0.96	35	8.0	0.21	40	11.5	0.20
-0.95	-0.76	17	8.4	0.24	21	11.5	0.22
-0.75	-0.56	25	8.5	0.26	29	11.7	0.23
-0.55	-0.36	48	8.0	0.25	37	12.0	0.23
-0.35	-0.16	38	8.3	0.24	34	11.6	0.22
-0.15	0.04	53	8.4	0.24	62	10.8	0.22
0.05	0.24	35	7.8	0.23	25	11.1	0.22
0.25	0.44	30	7.4	0.25	48	10.7	0.24
0.45	0.64	43	7.3	0.25	24	10.7	0.22
0.65	0.84	19	7.5	0.25	20	11.0	0.21
0.85	1.04	11	8.3	0.21	15	11.2	0.25
1.05	1.24	18	8.0	0.24	16	11.3	0.24
1.25	1.44	10	8.0	0.17	14	11.1	0.21
1.45	1.64	2	8.0	0.42	1	11.0	0.21
1.65	1.84	2	6.0	0.24	4	11.0	0.22
1.85	2.04	14	6.0	0.25	3	11.0	0.23
2.05	2.24	0	0.0	0.0	8	9.0	0.24
2.25	2.44	0	0.0	0.0	0	0.0	0.0
2.45	2.64	0	0.0	0.0	0	0.0	0.0

Table 7.1 Mean Number of Items and Posterior Variances for Subtest 1 Under Criteria Levels 0.10 and 0.05

SUBTEST 1							
		Criterion = 0.025			Criterion = 0.01		
θ Interval		f	n of Items	Post. Var.	f	n of Items	Post. Var.
-3.55	-3.36	0	0.0	0.0	0	0.0	0.0
-3.35	-3.16	0	0.0	0.0	0	0.0	0.0
-3.15	-2.96	0	0.0	0.0	0	0.0	0.0
-2.95	-2.76	0	0.0	0.0	0	0.0	0.0
-2.75	-2.56	0	0.0	0.0	0	0.0	0.0
-2.55	-2.36	0	0.0	0.0	0	0.0	0.0
-2.35	-2.16	0	0.0	0.0	1	13.0	0.17
-2.15	-1.96	4	11.0	0.17	4	13.0	0.17
-1.95	-1.76	11	11.5	0.18	8	13.3	0.17
-1.75	-1.56	6	12.2	0.17	9	15.1	0.18
-1.55	-1.36	17	12.4	0.17	16	15.4	0.17
-1.35	-1.16	16	12.4	0.20	19	15.3	0.20
-1.15	-0.96	35	12.6	0.19	33	15.5	0.19
-0.95	-0.76	26	13.2	0.21	20	15.7	0.19
-0.75	-0.56	31	13.8	0.22	35	15.7	0.22
-0.55	-0.36	37	14.2	0.22	35	16.2	0.22
-0.35	-0.16	42	14.3	0.23	51	16.5	0.22
-0.15	0.04	49	14.9	0.21	36	17.0	0.20
0.05	0.24	32	14.6	0.22	34	17.1	0.22
0.25	0.44	38	12.8	0.23	44	17.2	0.22
0.45	0.64	28	13.3	0.20	30	16.7	0.21
0.65	0.84	21	14.1	0.21	16	17.3	0.20
0.85	1.04	12	13.4	0.23	12	16.3	0.28
1.05	1.24	17	13.8	0.24	20	16.2	0.21
1.25	1.44	14	13.1	0.22	13	15.8	0.23
1.45	1.64	3	12.7	0.20	4	15.0	0.20
1.65	1.84	3	12.0	0.22	3	15.0	0.22
1.85	2.04	6	12.0	0.23	4	15.0	0.23
2.05	2.24	5	12.0	0.24	6	14.2	0.24
2.25	2.44	0	0.0	0.0	0	0.0	0.0
2.45	2.64	0	0.0	0.0	0	0.0	0.0

Table 7.2 Mean Number of Items and Posterior Variances for Subtest 1 Under Criteria Levels 0.025 and 0.01

SUBTEST 1							
		Criterion = 0.001			Criterion = 0.0		
Interval		f	n of Items	Post. Var.	f	n of Items	Post. Var.
-3.55	-3.36	0	0.0	0.0	0	0.0	0.0
-3.35	-3.16	0	0.0	0.0	0	0.0	0.0
-3.15	-2.96	0	0.0	0.0	0	0.0	0.0
-2.95	-2.76	0	0.0	0.0	0	0.0	0.0
-2.75	-2.56	0	0.0	0.0	0	0.0	0.0
-2.55	-2.36	0	0.0	0.0	0	0.0	0.0
-2.35	-2.16	1	16.0	0.17	1	20.0	0.17
-2.15	-1.96	4	16.0	0.17	4	20.0	0.17
-1.95	-1.76	8	16.1	0.17	8	20.0	0.17
-1.75	-1.56	10	16.2	0.18	10	20.0	0.18
-1.55	-1.36	17	16.3	0.17	16	20.0	0.17
-1.35	-1.16	14	16.4	0.19	15	20.0	0.20
-1.15	-0.96	37	16.5	0.20	34	20.0	0.19
-0.95	-0.76	26	16.8	0.20	28	20.0	0.21
-0.75	-0.56	27	17.2	0.21	26	20.0	0.21
-0.55	-0.36	37	17.2	0.22	37	20.0	0.21
-0.35	-0.16	43	17.8	0.21	44	20.0	0.21
-0.15	0.04	44	18.0	0.21	43	20.0	0.22
0.05	0.24	37	18.5	0.22	36	20.0	0.22
0.25	0.44	40	19.1	0.21	43	20.0	0.22
0.45	0.64	29	19.6	0.21	29	20.0	0.21
0.65	0.84	18	19.3	0.22	18	20.0	0.22
0.85	1.04	10	19.3	0.23	10	20.0	0.23
1.05	1.24	21	19.2	0.22	21	20.0	0.22
1.25	1.44	11	18.2	0.22	11	20.0	0.22
1.45	1.64	6	18.0	0.22	6	20.0	0.22
1.65	1.84	0	0.0	0.0	0	0.0	0.0
1.85	2.04	7	18.0	0.23	7	20.0	0.23
2.05	2.24	6	17.0	0.23	6	20.0	0.23
2.25	2.44	0	0.0	0.0	0	0.0	0.0
2.45	2.64	0	0.0	0.0	0	0.0	0.0

Table 7.3 Mean Number of Items and Posterior Variances for Subtest 1 Under Criteria Levels 0.001 and 0.0

SUBTEST 2							
Θ Interval		Criterion = 0.10			Criterion = 0.05		
		f	n of Items	Post. Var.	f	n of Items	Post. Var.
-3.55	-3.36	0	0.0	0.0	0	0.0	0.0
-3.35	-3.16	0	0.0	0.0	0	0.0	0.0
-3.15	-2.96	0	0.0	0.0	0	0.0	0.0
-2.95	-2.76	0	0.0	0.0	1	15.0	0.19
-2.75	-2.56	1	11.0	0.21	0	0.0	0.0
-2.55	-2.36	0	0.0	0.0	2	16.5	0.17
-2.35	-2.16	3	12.0	0.19	2	16.5	0.19
-2.15	-1.96	4	11.5	0.20	2	17.0	0.18
-1.95	-1.76	8	12.5	0.18	10	17.4	0.16
-1.75	-1.56	10	13.8	0.17	9	18.3	0.16
-1.55	-1.36	15	13.9	0.16	13	18.8	0.16
-1.35	-1.16	15	14.6	0.16	19	19.2	0.14
-1.15	-0.96	26	15.5	0.15	21	19.3	0.15
-0.95	-0.76	20	15.2	0.15	27	19.9	0.14
-0.75	-0.56	43	14.1	0.14	43	20.0	0.14
-0.55	-0.36	38	13.0	0.15	37	20.0	0.14
-0.35	-0.16	47	13.0	0.15	39	20.0	0.14
-0.15	0.04	41	12.1	0.17	47	19.0	0.15
0.05	0.24	31	11.5	0.17	32	16.0	0.17
0.25	0.44	23	10.1	0.19	26	15.4	0.18
0.45	0.64	12	8.8	0.23	16	13.4	0.21
0.65	0.84	34	7.4	0.25	18	12.9	0.22
0.85	1.04	41	6.6	0.26	36	11.6	0.24
1.05	1.24	27	5.2	0.32	32	10.7	0.26
1.25	1.44	7	4.0	0.40	8	8.8	0.29
1.45	1.64	0	0.0	0.0	1	7.0	0.41
1.65	1.84	1	4.0	0.46	5	8.0	0.41
1.85	2.04	6	3.0	0.50	7	3.3	0.51
2.05	2.24	0	0.0	0.0	0	0.0	0.0
2.25	2.44	0	0.0	0.0	0	0.0	0.0
2.45	2.64	0	0.0	0.0	0	0.0	0.0

Table 7.4 Mean Number of Items and Posterior Variances for Subtest 2 Under Criteria Levels 0.10 and 0.05

SUBTEST 2							
		Criterion = 0.025			Criterion = 0.01		
Θ Interval		f	n of Items	Post. Var.	f	n of Items	Post. Var.
-3.55	-3.36	0	0.0	0.0	0	0.0	0.0
-3.35	-3.16	0	0.0	0.0	0	0.0	0.0
-3.15	-2.96	1	15.0	0.19	1	16.0	0.19
-2.95	-2.76	0	0.0	0.0	0	0.0	0.0
-2.75	-2.56	0	0.0	0.0	0	0.0	0.0
-2.55	-2.36	2	17.0	0.17	2	17.0	0.17
-2.35	-2.16	2	16.5	0.19	2	17.5	0.18
-2.15	-1.96	2	17.0	0.18	3	18.3	0.16
-1.95	-1.76	10	18.4	0.16	9	18.3	0.16
-1.75	-1.56	9	18.7	0.16	9	18.7	0.16
-1.55	-1.36	13	18.8	0.16	13	18.9	0.15
-1.35	-1.16	20	19.3	0.14	20	19.1	0.15
-1.15	-0.96	21	19.3	0.15	20	19.7	0.15
-0.95	-0.76	25	20.0	0.14	26	20.0	0.14
-0.75	-0.56	46	20.0	0.14	47	20.0	0.14
-0.55	-0.36	36	20.0	0.14	35	20.0	0.14
-0.35	-0.16	39	20.0	0.14	39	20.0	0.14
-0.15	0.04	48	20.0	0.15	48	20.0	0.15
0.05	0.24	24	20.0	0.16	25	20.0	0.16
0.25	0.44	34	20.0	0.17	33	20.0	0.17
0.45	0.64	13	19.2	0.20	16	20.0	0.20
0.65	0.84	24	16.6	0.22	20	20.0	0.21
0.85	1.04	33	16.1	0.23	33	20.0	0.23
1.05	1.24	32	15.2	0.25	31	18.2	0.24
1.25	1.44	5	13.6	0.27	5	16.2	0.27
1.45	1.64	2	12.0	0.37	3	16.3	0.35
1.65	1.84	5	10.6	0.39	4	15.8	0.38
1.85	2.04	5	7.8	0.47	5	14.0	0.42
2.05	2.24	2	7.0	0.48	4	12.5	0.45
2.25	2.44	0	0.0	0.0	0	0.0	0.0
2.45	2.64	0	0.0	0.0	0	0.0	0.0

Table 7.5 Mean Number of Items and Posterior Variances for Subtest 2 Under Criteria Levels 0.025 and 0.01

SUBTEST 2							
		Criterion = 0.001			Criterion = 0.0		
Θ Interval		f	n of Items	Post. Var.	f	n of Items	Post. Var.
-3.55	-3.36	0	0.0	0.0	0	0.0	0.0
-3.35	-3.16	0	0.0	0.0	0	0.0	0.0
-3.15	-2.96	1	17.0	0.19	1	21.0	0.19
-2.95	-2.76	0	0.0	0.0	0	0.0	0.0
-2.75	-2.56	0	0.0	0.0	0	0.0	0.0
-2.55	-2.36	2	18.0	0.17	2	21.0	0.17
-2.35	-2.16	2	18.0	0.18	2	21.0	0.18
-2.15	-1.96	2	18.0	0.18	2	21.0	0.18
-1.95	-1.76	10	18.4	0.16	10	21.0	0.16
-1.75	-1.56	9	18.7	0.16	9	21.0	0.16
-1.55	-1.36	13	18.8	0.16	14	21.0	0.15
-1.35	-1.16	20	19.4	0.15	18	21.0	0.14
-1.15	-0.96	19	20.0	0.14	20	21.0	0.14
-0.95	-0.76	26	20.0	0.14	26	21.0	0.14
-0.75	-0.56	48	20.0	0.14	49	21.0	0.14
-0.55	-0.36	35	20.0	0.14	34	21.0	0.14
-0.35	-0.16	39	20.0	0.14	40	21.0	0.14
-0.15	0.04	48	20.0	0.15	47	21.0	0.15
0.05	0.24	25	20.0	0.16	25	21.0	0.16
0.25	0.44	33	20.0	0.18	32	21.0	0.18
0.45	0.64	17	20.1	0.20	18	21.0	0.21
0.65	0.84	21	20.0	0.21	18	21.0	0.21
0.85	1.04	34	20.2	0.23	36	21.0	0.21
1.05	1.24	24	20.8	0.21	24	21.0	0.23
1.25	1.44	9	21.0	0.29	10	21.0	0.29
1.45	1.64	4	20.5	0.34	4	21.0	0.33
1.65	1.84	4	19.0	0.38	4	21.0	0.37
1.85	2.04	3	18.7	0.38	3	21.0	0.38
2.05	2.24	5	17.0	0.45	5	21.0	0.43
2.25	2.44	0	0.0	0.0	0	0.0	0.0
2.45	2.64	0	0.0	0.0	0	0.0	0.0

Table 7.6 Mean Number of Items and Posterior Variances for Subtest 2 Under Criteria Levels 0.001 and 0.0

SUBTEST 3							
		Criterion = 0.10			Criterion = 0.05		
Θ Interval		f	n of Items	Post. Var.	f	n of Items	Post. Var.
-3.55	-3.36	0	0.0	0.0	0	0.0	0.0
-3.35	-3.16	0	0.0	0.0	0	0.0	0.0
-3.15	-2.96	0	0.0	0.0	1	9.0	0.30
-2.95	-2.76	0	0.0	0.0	0	0.0	0.0
-2.75	-2.56	0	0.0	0.0	1	10.0	0.31
-2.55	-2.36	4	4.8	0.35	2	11.0	0.30
-2.35	-2.16	2	6.0	0.35	2	12.0	0.30
-2.15	-1.96	3	9.7	0.26	2	15.0	0.20
-1.95	-1.76	4	11.0	0.27	5	14.4	0.25
-1.75	-1.56	10	11.3	0.24	13	14.8	0.23
-1.55	-1.36	17	11.2	0.21	16	16.7	0.19
-1.35	-1.16	12	11.1	0.23	19	18.9	0.19
-1.15	-0.96	46	11.2	0.21	37	19.5	0.18
-0.95	-0.76	35	12.6	0.19	37	19.8	0.16
-0.75	-0.56	43	13.6	0.18	40	20.0	0.17
-0.55	-0.36	33	13.0	0.17	34	20.0	0.17
-0.35	-0.16	36	12.9	0.18	37	18.8	0.16
-0.15	0.04	34	11.6	0.17	36	18.1	0.17
0.05	0.24	30	11.0	0.20	26	18.1	0.18
0.25	0.44	26	10.9	0.22	31	18.0	0.19
0.45	0.64	18	10.1	0.24	17	18.3	0.19
0.65	0.84	27	8.8	0.22	25	18.0	0.20
0.85	1.04	27	7.4	0.23	19	18.0	0.19
1.05	1.24	14	7.3	0.26	21	17.4	0.21
1.25	1.44	10	7.3	0.31	4	17.3	0.25
1.45	1.64	5	8.4	0.30	9	16.8	0.25
1.65	1.84	7	7.6	0.32	12	13.1	0.29
1.85	2.04	7	6.7	0.34	4	12.0	0.31
2.05	2.24	1	5.0	0.43	2	11.5	0.31
2.25	2.44	2	5.0	0.35	1	8.0	0.34
2.45	2.64	0	0.0	0.0	0	0.0	0.0

Table 7.7 Mean Number of Items and Posterior Variances for Subtest 3 Under Criteria Levels 0.10 and 0.05

SUBTEST 3							
		Criterion = 0.025			Criterion = 0.01		
Θ Interval		f	n of Items	Post. Var.	f	n of Items	Post. Var.
-3.55	-3.36	0	0.0	0.0	0	0.0	0.0
-3.35	-3.16	1	13.0	0.28	1	15.0	0.27
-3.15	-2.96	0	0.0	0.0	0	0.0	0.0
-2.95	-2.76	0	0.0	0.0	0	0.0	0.0
-2.75	-2.56	2	14.5	0.28	2	18.0	0.27
-2.55	-2.36	2	15.0	0.27	3	18.0	0.28
-2.35	-2.16	2	17.5	0.25	1	20.0	0.21
-2.15	-1.96	1	18.0	0.19	1	20.0	0.19
-1.95	-1.76	3	18.0	0.27	5	20.4	0.24
-1.75	-1.56	16	20.2	0.21	14	20.9	0.21
-1.55	-1.36	16	21.2	0.19	17	21.2	0.19
-1.35	-1.16	19	21.3	0.19	18	21.3	0.19
-1.15	-0.96	40	21.6	0.18	39	22.2	0.18
-0.95	-0.76	28	22.0	0.16	28	23.0	0.16
-0.75	-0.56	46	22.0	0.17	46	23.0	0.17
-0.55	-0.36	35	22.1	0.16	35	23.0	0.16
-0.35	-0.16	34	23.1	0.16	35	23.1	0.16
-0.15	0.04	38	23.1	0.16	38	23.1	0.17
0.05	0.24	27	23.0	0.18	26	23.0	0.18
0.25	0.44	27	22.5	0.18	29	23.2	0.18
0.45	0.64	19	22.4	0.18	23	23.6	0.18
0.65	0.84	20	22.4	0.19	15	24.0	0.18
0.85	1.04	24	21.1	0.19	25	23.5	0.19
1.05	1.24	20	19.9	0.20	18	22.8	0.20
1.25	1.44	8	19.4	0.25	10	22.3	0.23
1.45	1.64	10	19.1	0.25	9	20.9	0.25
1.65	1.84	5	17.8	0.27	5	19.6	0.27
1.85	2.04	8	17.6	0.28	5	19.6	0.28
2.05	2.24	0	0.0	0.0	3	19.7	0.28
2.25	2.44	2	17.0	0.32	2	19.0	0.31
2.45	2.64	0	0.0	0.0	0	0.0	0.0

Table 7.8 Mean Number of Items and Posterior Variances for Subtest 3 Under Criteria Levels 0.025 and 0.01

SUBTEST 3							
		Criterion = 0.001			Criterion = 0.0		
Θ Interval		f	n of Items	Post. Var.	f	n of Items	Post. Var.
-3.55	-3.36	0	0.0	0.0	0	0.0	0.0
-3.35	-3.16	1	18.0	0.26	1	24.0	0.27
-3.15	-2.96	0	0.0	0.0	0	0.0	0.0
-2.95	-2.76	0	0.0	0.0	0	0.0	0.0
-2.75	-2.56	3	20.0	0.27	3	24.0	0.27
-2.55	-2.36	2	20.0	0.28	2	24.0	0.28
-2.35	-2.16	1	21.0	0.20	1	24.0	0.20
-2.15	-1.96	1	22.0	0.19	1	24.0	0.19
-1.95	-1.76	6	21.5	0.23	5	24.0	0.23
-1.75	-1.56	13	22.0	0.21	14	24.0	0.21
-1.55	-1.36	17	22.3	0.19	17	24.0	0.19
-1.35	-1.16	17	22.4	0.19	16	24.0	0.19
-1.15	-0.96	40	23.0	0.17	41	24.0	0.17
-0.95	-0.76	28	23.0	0.16	28	24.0	0.16
-0.75	-0.56	46	23.0	0.17	45	24.0	0.17
-0.55	-0.36	36	23.0	0.16	36	24.0	0.16
-0.35	-0.16	34	23.1	0.16	35	24.0	0.16
-0.15	0.04	37	23.1	0.16	37	24.0	0.16
0.05	0.24	27	23.0	0.18	26	24.0	0.17
0.25	0.44	29	23.1	0.18	30	24.0	0.18
0.45	0.64	20	23.9	0.17	19	24.0	0.17
0.65	0.84	22	24.0	0.20	23	24.0	0.20
0.85	1.04	21	24.0	0.18	21	24.0	0.18
1.05	1.24	19	24.0	0.20	19	24.0	0.20
1.25	1.44	9	24.0	0.23	9	24.0	0.23
1.45	1.64	9	23.1	0.25	8	24.0	0.24
1.65	1.84	4	22.5	0.26	5	24.0	0.26
1.85	2.04	6	22.7	0.28	6	24.0	0.28
2.05	2.24	3	22.3	0.30	3	24.0	0.30
2.25	2.44	2	20.5	0.31	2	24.0	0.31
2.45	2.64	0	0.0	0.0	0	0.0	0.0

Table 7.9 Mean Number of Items and Posterior Variances for Subtest 3 Under Criteria Levels 0.001 and 0.0

SUBTEST 4							
		Criterion = 0.10			Criterion = 0.05		
Θ Interval		f	n of Items	Post. Var.	f	n of Items	Post. Var.
-3.55	-3.36	0	0.0	0.0	0	0.0	0.0
-3.35	-3.16	0	0.0	0.0	0	0.0	0.0
-3.15	-2.96	0	0.0	0.0	0	0.0	0.0
-2.95	-2.76	2	4.5	0.23	2	9.5	0.20
-2.75	-2.56	0	0.0	0.0	0	0.0	0.0
-2.55	-2.36	1	6.0	0.16	3	11.3	0.22
-2.35	-2.16	3	7.3	0.21	2	12.0	0.18
-2.15	-1.96	3	7.7	0.21	2	12.0	0.19
-1.95	-1.76	5	8.0	0.19	9	12.9	0.19
-1.75	-1.56	17	8.0	0.20	14	13.6	0.19
-1.55	-1.36	18	7.4	0.21	21	14.6	0.20
-1.35	-1.16	21	7.6	0.22	24	15.5	0.20
-1.15	-0.96	14	8.6	0.21	15	15.7	0.20
-0.95	-0.76	16	7.8	0.23	22	16.1	0.18
-0.75	-0.56	28	6.8	0.23	26	16.1	0.20
-0.55	-0.36	46	6.0	0.23	36	16.2	0.20
-0.35	-0.16	54	6.0	0.23	41	16.0	0.20
-0.15	0.04	33	6.2	0.25	46	16.3	0.20
0.05	0.24	34	5.3	0.31	39	13.9	0.22
0.25	0.44	49	5.7	0.24	37	13.0	0.20
0.45	0.64	22	6.2	0.23	25	13.0	0.19
0.65	0.84	19	7.0	0.21	22	13.0	0.20
0.85	1.04	16	6.9	0.21	16	13.0	0.19
1.05	1.24	5	5.8	0.26	10	12.9	0.22
1.25	1.44	15	4.9	0.27	13	11.8	0.24
1.45	1.64	17	3.6	0.29	12	11.5	0.24
1.65	1.84	11	3.0	0.32	10	11.1	0.25
1.85	2.04	3	3.0	0.30	5	9.0	0.28
2.05	2.24	1	3.0	0.32	1	6.0	0.30
2.25	2.44	0	0.0	0.0	0	0.0	0.0
2.45	2.64	0	0.0	0.0	0	0.0	0.0

Table 7.10 Mean Number of Items and Posterior Variances for Subtest 4 Under Criteria Levels 0.10 and 0.05

SUBTEST 4							
		Criterion = 0.025			Criterion = 0.01		
θ Interval		f	n of Items	Post. Var.	f	n of Items	Post. Var.
-3.55	-3.36	0	0.0	0.0	0	0.0	0.0
-3.35	-3.16	0	0.0	0.0	0	0.0	0.0
-3.15	-2.96	1	13.0	0.23	1	15.0	0.23
-2.95	-2.76	1	15.0	0.16	1	16.0	0.16
-2.75	-2.56	0	0.0	0.0	1	15.0	0.24
-2.55	-2.36	3	14.3	0.21	2	16.0	0.18
-2.35	-2.16	2	15.5	0.18	2	15.5	0.20
-2.15	-1.96	4	16.0	0.18	4	17.0	0.18
-1.95	-1.76	5	16.0	0.19	5	17.0	0.20
-1.75	-1.56	19	16.1	0.19	18	17.0	0.19
-1.55	-1.36	20	16.1	0.19	20	17.1	0.18
-1.35	-1.16	21	17.2	0.20	22	17.2	0.20
-1.15	-0.96	16	17.3	0.19	16	17.6	0.19
-0.95	-0.76	21	17.0	0.19	24	18.8	0.19
-0.75	-0.56	26	17.0	0.20	26	18.1	0.19
-0.55	-0.36	35	17.5	0.19	34	18.0	0.19
-0.35	-0.16	50	17.7	0.20	50	18.1	0.20
-0.15	0.04	45	17.4	0.20	43	18.4	0.20
0.05	0.24	37	17.8	0.20	36	18.4	0.19
0.25	0.44	34	18.1	0.19	36	18.1	0.20
0.45	0.64	29	18.0	0.20	29	18.1	0.20
0.65	0.84	21	16.1	0.19	21	18.0	0.18
0.85	1.04	12	16.1	0.20	11	18.1	0.20
1.05	1.24	12	16.0	0.21	15	18.0	0.21
1.25	1.44	15	15.0	0.23	12	16.3	0.22
1.45	1.64	9	14.6	0.24	9	15.6	0.24
1.65	1.84	11	14.2	0.25	12	15.2	0.25
1.85	2.04	4	14.0	0.28	3	15.0	0.28
2.05	2.24	0	0.0	0.0	0	0.0	0.0
2.25	2.44	0	0.0	0.0	0	0.0	0.0
2.45	2.64	0	0.0	0.0	0	0.0	0.0

Table 7.11 Mean Number of Items and Posterior Variances for Subtest 4 Under Criteria Levels 0.025 and 0.01

SUBTEST 4							
		Criterion = 0.001			Criterion = 0.0		
Θ Interval		f	n of Items	Post. Var.	f	n of Items	Post. Var.
-3.55	-3.36	0	0.0	0.0	0	0.0	0.0
-3.35	-3.16	0	0.0	0.0	0	0.0	0.0
-3.15	-2.96	1	16.0	0.23	1	21.0	0.24
-2.95	-2.76	1	17.0	0.16	1	21.0	0.16
-2.75	-2.56	1	16.0	0.24	1	21.0	0.23
-2.55	-2.36	2	17.0	0.18	2	21.0	0.18
-2.35	-2.16	2	16.5	0.20	2	21.0	0.23
-2.15	-1.96	4	17.0	0.18	4	21.0	0.18
-1.95	-1.76	5	17.2	0.20	6	21.0	0.19
-1.75	-1.56	18	18.0	0.19	17	21.0	0.19
-1.55	-1.36	21	18.1	0.18	20	21.0	0.18
-1.35	-1.16	21	18.2	0.20	22	21.0	0.20
-1.15	-0.96	18	18.6	0.19	18	21.0	0.19
-0.95	-0.76	21	19.0	0.18	22	21.0	0.19
-0.75	-0.56	28	19.1	0.19	31	21.0	0.19
-0.55	-0.36	35	19.1	0.19	34	21.0	0.19
-0.35	-0.16	51	18.9	0.20	50	21.0	0.20
-0.15	0.04	44	18.8	0.19	40	21.0	0.19
0.05	0.24	30	20.0	0.19	35	21.0	0.19
0.25	0.44	38	20.0	0.20	38	21.0	0.20
0.45	0.64	29	19.2	0.20	27	21.0	0.20
0.65	0.84	21	18.1	0.18	21	21.0	0.18
0.85	1.04	11	18.1	0.20	11	21.0	0.20
1.05	1.24	16	18.0	0.21	15	21.0	0.21
1.25	1.44	10	18.0	0.22	10	21.0	0.22
1.45	1.64	10	17.9	0.23	10	21.0	0.23
1.65	1.84	11	17.2	0.25	11	21.0	0.24
1.85	2.04	4	17.0	0.27	4	21.0	0.27
2.05	2.24	0	0.0	0.0	0	0.0	0.0
2.25	2.44	0	0.0	0.0	0	0.0	0.0
2.45	2.64	0	0.0	0.0	0	0.0	0.0

Table 7.12 Mean Number of Items and Posterior Variances for Subtest 4 Under Criteria Levels 0.001 and 0.0

8. Appendix C

The following tables contain the observed efficiency points for each Θ interval from the simulated testings. This data is graphically displayed in Figures 4.15 through 4.18.

Observed Efficiencies For Subtest 1

Interval Range		C_{10}	C_{05}	C_{025}	C_{01}	C_{001}
-3.55	-3.36	0.0	0.0	0.0	0.0	0.0
-3.35	-3.16	0.0	0.0	0.0	0.0	0.0
-3.15	-2.96	0.0	0.0	0.0	0.0	0.0
-2.95	-2.76	0.0	0.0	0.0	0.0	0.0
-2.75	-2.56	0.0	0.0	0.0	0.0	0.0
-2.55	-2.36	0.0	0.0	0.0	0.0	0.0
-2.35	-2.16	0.0	0.0	0.0	0.98	1.00
-2.15	-1.96	0.82	0.94	0.95	0.99	1.00
-1.95	-1.76	0.89	0.96	0.98	0.99	1.00
-1.75	-1.56	0.86	0.95	0.93	0.98	0.99
-1.55	-1.36	0.88	0.94	0.95	0.96	1.00
-1.35	-1.16	0.92	0.94	0.98	1.00	1.00
-1.15	-0.96	0.92	0.98	0.98	1.00	1.00
-0.95	-0.76	0.90	0.97	0.99	1.00	0.99
-0.75	-0.56	0.91	0.95	0.97	1.01	0.99
-0.55	-0.36	0.86	0.94	0.96	0.97	0.98
-0.35	-0.16	0.85	0.94	0.98	0.99	1.00
-0.15	0.04	0.86	0.91	0.97	0.99	0.99
0.05	0.24	0.84	0.97	1.01	1.01	1.01
0.25	0.44	0.82	0.96	1.00	0.98	1.00
0.45	0.64	0.90	0.94	0.93	0.99	0.99
0.65	0.84	0.92	0.97	0.98	0.99	0.99
0.85	1.04	0.94	0.98	0.99	0.99	1.00
1.05	1.24	0.94	0.98	0.99	0.99	1.00
1.25	1.44	0.95	0.98	0.99	0.99	0.99
1.45	1.64	0.89	0.98	0.98	0.98	0.99
1.65	1.84	0.0	0.0	0.0	0.0	0.0
1.85	2.04	0.82	0.92	0.93	0.94	1.00
2.05	2.24	0.0	1.11	1.02	0.99	1.00
2.25	2.44	0.0	0.0	0.0	0.0	0.0
2.45	2.64	0.0	0.0	0.0	0.0	0.0

Table 8.1 Observed Efficiencies For Subtest 1

Observed Efficiencies For Subtest 2

Interval Range		C_{10}	C_{05}	C_{025}	C_{01}	C_{001}
-3.55	-3.36	0.0	0.0	0.0	0.0	0.0
-3.35	-3.16	0.0	0.0	0.0	0.0	0.0
-3.15	-2.96	0.0	0.0	1.01	1.00	1.00
-2.95	-2.76	0.0	0.0	0.0	0.0	0.0
-2.75	-2.56	0.0	0.0	0.0	0.0	0.0
-2.55	-2.36	0.0	0.98	0.99	0.98	0.99
-2.35	-2.16	0.89	1.00	1.00	1.00	1.00
-2.15	-1.96	0.84	0.99	0.99	1.00	1.00
-1.95	-1.76	0.88	0.99	1.00	1.00	0.99
-1.75	-1.56	0.90	0.99	0.99	0.99	0.99
-1.55	-1.36	0.89	0.99	0.99	0.99	0.99
-1.35	-1.16	0.90	1.00	1.00	1.00	1.00
-1.15	-0.96	0.92	0.99	0.99	0.99	1.00
-0.95	-0.76	0.91	0.99	1.00	1.00	1.00
-0.75	-0.56	0.88	0.99	0.99	1.00	0.99
-0.55	-0.36	0.91	0.98	0.99	0.99	0.99
-0.35	-0.16	0.94	0.99	0.99	0.99	0.99
-0.15	0.04	0.94	0.99	0.99	0.99	0.99
0.05	0.24	0.94	0.97	1.00	1.00	0.99
0.25	0.44	0.95	0.99	1.01	1.00	1.00
0.45	0.64	0.87	0.97	0.98	1.00	0.99
0.65	0.84	0.80	0.92	0.98	1.00	0.99
0.85	1.04	0.71	0.90	0.94	0.96	0.95
1.05	1.24	0.56	0.83	0.92	0.96	1.03
1.25	1.44	0.47	0.72	0.88	0.94	0.99
1.45	1.64	0.0	0.69	1.10	1.01	0.98
1.65	1.84	1.11	0.97	0.98	0.95	0.97
1.85	2.04	1.00	0.97	1.02	1.00	0.99
2.05	2.24	0.0	0.0	1.00	0.99	1.01
2.25	2.44	0.0	0.0	0.0	0.0	0.0
2.45	2.64	0.0	0.0	0.0	0.0	0.0

Table 8.2 Observed Efficiencies For Subtest 2

Observed Efficiencies For Subtest 3						
Interval Range		C_{10}	C_{05}	C_{025}	C_{01}	C_{001}
-3.55	-3.36	0.0	0.0	0.0	0.0	0.0
-3.35	-3.16	0.0	0.0	0.98	0.99	0.99
-3.15	-2.96	0.0	0.0	0.0	0.0	0.0
-2.95	-2.76	0.0	0.0	0.0	0.0	0.0
-2.75	-2.56	0.0	0.82	0.92	0.99	1.00
-2.55	-2.36	0.51	0.82	0.93	0.98	0.99
-2.35	-2.16	0.59	0.84	0.95	1.00	1.00
-2.15	-1.96	0.75	0.91	0.96	1.00	0.99
-1.95	-1.76	0.80	0.91	0.96	0.99	0.99
-1.75	-1.56	0.79	0.90	0.98	0.99	1.00
-1.55	-1.36	0.75	0.92	0.99	0.99	1.00
-1.35	-1.16	0.73	0.95	0.99	0.99	0.99
-1.15	-0.96	0.75	0.98	1.00	0.99	1.00
-0.95	-0.76	0.80	0.95	0.98	0.99	0.99
-0.75	-0.56	0.88	1.01	1.00	1.00	1.00
-0.55	-0.36	0.89	0.97	0.99	0.99	1.00
-0.35	-0.16	0.90	0.98	1.00	1.00	1.00
-0.15	0.04	0.90	0.97	1.00	0.99	1.00
0.05	0.24	0.87	0.95	0.99	0.99	0.99
0.25	0.44	0.85	0.97	0.99	0.99	0.99
0.45	0.64	0.79	0.97	1.00	0.98	1.00
0.65	0.84	0.71	0.94	1.02	1.00	0.99
0.85	1.04	0.65	0.92	0.98	0.99	0.99
1.05	1.24	0.75	0.94	0.98	1.01	0.99
1.25	1.44	0.84	0.94	1.02	0.99	0.99
1.45	1.64	0.88	0.96	0.98	0.99	0.99
1.65	1.84	0.90	0.94	0.94	0.96	0.98
1.85	2.04	0.92	0.98	0.97	1.01	1.02
2.05	2.24	0.82	0.78	0.0	1.06	1.00
2.25	2.44	0.87	0.83	1.05	0.99	1.00
2.45	2.64	0.0	0.0	0.0	0.0	0.0

Table 8.3 Observed Efficiencies For Subtest 3

Observed Efficiencies For Subtest 4						
Interval Range		C_{10}	C_{05}	C_{025}	C_{01}	C_{001}
-3.55	-3.36	0.0	0.0	0.0	0.0	0.0
-3.35	-3.16	0.0	0.0	0.0	0.0	0.0
-3.15	-2.96	0.0	0.0	0.97	1.00	0.99
-2.95	-2.76	0.75	0.91	0.98	0.99	0.99
-2.75	-2.56	0.0	0.0	0.0	0.99	1.00
-2.55	-2.36	0.71	0.92	0.98	0.99	1.00
-2.35	-2.16	0.75	0.91	0.99	0.99	1.00
-2.15	-1.96	0.74	0.91	0.99	0.99	0.99
-1.95	-1.76	0.74	0.91	0.98	0.99	0.99
-1.75	-1.56	0.80	0.97	1.00	0.99	0.99
-1.55	-1.36	0.83	0.99	1.02	1.00	1.00
-1.35	-1.16	0.85	0.97	0.99	0.99	0.99
-1.15	-0.96	0.89	0.98	0.99	0.99	0.99
-0.95	-0.76	0.84	0.98	1.00	1.00	0.99
-0.75	-0.56	0.81	0.97	1.00	1.00	1.00
-0.55	-0.36	0.76	0.98	0.99	1.00	1.00
-0.35	-0.16	0.79	0.98	0.99	0.99	1.00
-0.15	0.04	0.84	0.98	1.00	1.00	1.00
0.05	0.24	0.84	0.95	0.99	0.99	1.00
0.25	0.44	0.84	0.96	0.99	0.99	1.00
0.45	0.64	0.83	0.95	0.99	0.99	0.99
0.65	0.84	0.87	0.94	1.00	0.98	0.98
0.85	1.04	0.88	0.97	1.02	1.00	1.00
1.05	1.24	0.91	0.98	0.99	1.00	1.00
1.25	1.44	0.92	1.00	0.98	1.00	1.00
1.45	1.64	0.83	0.93	0.96	0.98	0.99
1.65	1.84	0.82	1.09	1.02	0.98	1.00
1.85	2.04	0.74	0.86	1.03	0.96	1.02
2.05	2.24	0.0	0.0	0.0	0.0	0.0
2.25	2.44	0.0	0.0	0.0	0.0	0.0
2.45	2.64	0.0	0.0	0.0	0.0	0.0

Table 8.4 Observed Efficiencies For Subtest 4